BIO METRIA

III Encontro Luso-Galaico de Biometria

Livro de Atas

Departamento de Matemática
Universidade de Aveiro

28 - 30 junho 2018

Sociedade Portuguesa de Estatística

#SGaPEIO
Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións

Livro de Atas

III Encontro Luso-Galaico de Biometria

DMat, Univ. Aveiro

28 – 30 junho 2018

Promotores:

Sociedade Portuguesa de Estatística (SPE)

Sociedade Galega para a Promoción da Estatística e Investigación de Operacións (SGaPEIO)

Apoios:

- Departamento de Matemática, Universidade de Aveiro (DMat-UA)
- Fundação para a Ciência e a Tecnologia (FCT)
- Centro de Investigação e Desenvolvimento em Matemática e Aplicações, Universidade de Aveiro (CIDMA)[1]
- Linha Temática BioMath, CIDMA, Universidade de Aveiro
- Fábrica Centro Ciência Viva de Aveiro
- Centro de Estatística e Aplicações, Universidade de Lisboa (CEAUL)[1]
- Centro de Investigação em Tecnologias e Serviços de Saúde, Universidade do Porto (CINTESIS)
- Produtos e Serviços de Estatística, Lda (PSE)
- Instituto Nacional de Estatística (INE)
- Profijardim - Construção e Manutenção de Espaços Verdes, Lda (Profijardim)
- Administração do Porto de Aveiro, S.A. (APA, S.A.)
- Delta Cafés
- Edições Sílabo, Lda (eS)
- Jerónimo Martins, SGPS, SA (JM)

[1]O encontro e o presente livro de atas são parcialmente suportado por fundos portugueses através do Centro de Investigação e Desenvolvimento em Matemática e Aplicações, CIDMA, dentro do projeto UID/MAT/04106/2013 e do Centro de Estatística e Aplicações da Universidade de Lisboa, CEAUL, dentro do projeto UID/MAT/0006/2013.

Editores: Magda Monteiro, Adelaide Freitas, Laetitia Teixeira, Marco Costa

Título: Livro de Atas do III Encontro Luso-Galaico de Biometria

Editora: Sociedade Portuguesa de Estatística

Conceção Gráfica do Logotipo: Carina Sousa

ISBN: 978-972-8890-42-1

Apresentação

A cidade de Aveiro foi escolhida, pelas Sociedade Portuguesa de Estatística (SPE) e Sociedade Galega para a Promoción da Estatística e Investigación de Operacións (SGAPEIO), para acolher o III Encontro Luso-Galaico de Biometria (EBio2018).

Neste terceiro encontro sobre *Biometria*, realizado no Departamento de Matemática da Universidade de Aveiro, de 28 a 30 de junho de 2018, reunimos cerca de 100 participantes que desenvolvem e/ou aplicam metodologias estatísticas a dados das ciências da vida e do meio ambiente. Com o objetivo de criar sinergias entre diferentes áreas de aplicação e discutir os mais recentes desenvolvimentos metodológicos em Biometria, este encontro encoraja futuras colaborações e discussão entre os participantes.

O programa científico do EBio2018 é vasto e diversificado: composto por um minicurso intitulado "*Biometry with compositional data*", ministrado por Karel Hron; quatro sessões plenárias, com os oradores convidados Alessandro Fassò (Universidade de Bérgamo), Carlos Daniel Paulino (IST- Universidade de Lisboa), Peter Müller (Universidade do Texas Austin) e Ricardo Cao (Universidade da Corunha); oito sessões convidadas, com os oradores convidados Bruno Falissard (Universidade de Paris XI), Daniel Farewell (Universidade de Cardiff), Elizabeth Juarez-Colunga (Universidade de Colorado), María Xosé Rodríguez Álvarez (Centro Basco de Matemática Aplicada), Luiz Alexandre Peternelli (Universidade Federal de Viçosa), Lurdes Inoue (Universidade de Washington), Raquel Meneses (Universidade do Minho) e Ruwanthi Kolamunnage-Dona (Universidade de Liverpool); uma mesa-redonda, sobre "O papel do estatístico nas várias fases do ensaio clínico", que conta com o contributo de Aurora Baluja (Universidade de Santiago de Compostela), Elsa Branco (Novartis-Portugal), João Branco (IST- Universidade de Lisboa) e Júlio da Motta Singer (Universidade de São Paulo); 39 comunicações orais, divididas por 11 sessões abrangendo várias áreas do amplo campo da Biometria e, ainda, duas sessões de pósteres envolvendo 27 trabalhos. Agradecemos a todos o contributo para o sucesso científico deste encontro.

A Comissão Organizadora agradece ainda aos colegas da Comissão Científica, aos estudantes voluntários no apoio logístico e ainda às instituições que contribuíram para enriquecer este encontro com o seu apoio e dedicação. Um especial e profundo agradecimento aos parceiros e patrocinadores, mencionados na contracapa, que se associaram a este evento possibilitando não só a sua realização como também a atribuição de prémios à melhor comunicação oral apresentada por jovens investigadores e à melhor comunicação em formato de póster.

Em articulação com a Comissão Científica e os autores dos trabalhos aceites para o III Encontro Luso-Galaico de Biometria foi possível a elaboração do presente Livro de Atas no qual se disseminam, na forma de resumos alargados, a produção científica apresentada no evento. Esses resumos estão organizados neste documento de acordo com as sessões do programa científico (Minicurso, Sessões Plenárias, Sessões Convidadas, Sessões Prémio à Melhor Comunicação Oral apresentada por jovens investigadores e sessões com as contribuições dos Autores, em formato Oral e em formato Póster). A descrição das Comissões Organizadora

e Científica do EBio2018, do Programa Geral e do Programa Científico do encontro antecede os resumos das comunicações.

As publicações *online* do Guia do encontro e das Atas, com os resumos alargados das comunicações apresentadas, estão disponíveis no *site* do encontro (`http://ebio2018-pt.weebly.com/`).

Que este III Encontro Luso-Galaico de Biometria seja um espaço de excelentes dias de trabalho profícuo e convívio pelas terras dos marnotos!

Um bem-haja!

Aveiro, junho 2018

A Comissão Organizadora

Índice

COMISSÃO ORGANIZADORA

Magda Monteiro
Escola Superior de Tecnologia e Gestão de Águeda
Universidade de Aveiro, Portugal
Adelaide Freitas
Departamento de Matemática
Universidade de Aveiro, Portugal
Laetitia Teixeira
Instituto de Ciências Biomédicas Abel Salazar
Universidade do Porto, Portugal
Maria José Ginzo Villamayor
Departamento de Estatística, Análise Matemática e Investigação Operacional
Universidade de Santiago de Compostela, Espanha
Marco Costa
Escola Superior de Tecnologia e Gestão de Águeda
Universidade de Aveiro, Portugal
Paula Raña Míguez
Departamento de Matemáticas
Universidade da Corunha, Espanha

COMISSÃO CIENTÍFICA

Giovani Silva
Departamento de Matemática, Instituto Superior Técnico
Universidade de Lisboa, Portugal
Inês Sousa
Departamento de Matemática e Aplicações
Universidade do Minho, Portugal
Javier Roca Pardiñas
Departamento de Estatística e Investigación Operativa
Universidade de Vigo, Espanha
Lisete Sousa
Departamento de Estatística e Investigação Operacional
Universidade de Lisboa, Portugal
Maria Amalia Jácome Pumar
Departamento de Matemáticas
Universidade da Corunha, Espanha
Maria Teresa Seoane Pillado
Departamento de Ciências da Saúde
Universidade da Corunha, Espanha

Programa Geral

Quinta-feira, 28 de junho de 2018

8:30–9:00	Registo	sala: 11.1.28
9:00–11:00	Minicurso: ***Karel Hron*** *Biometry with compositional data*	Anf. 11.1.3
11:00–11:30	Pausa para Café	Bar Matemática
11:30–13:00	Minicurso (Cont.)	Anf. 11.1.3
13:00–14:30	Almoço	Cantina do Crasto
14:00–14:30	Registo	sala: 11.1.28
14:30–15:00	Sessão de Abertura	Anf. 11.1.3
15:00–16:00	Sessão Plenária I: ***Ricardo Cao*** *Nonparametric inference and covariate significance tests in mixture cure models*	Anf. 11.1.3
16:00–16:40	Pausa para Café	Sala de Professores, piso 3
16:00–16:40	Sessão de Pósteres I	Corredor, piso 3
16:40–18:00	Sessões Comunicações Orais I, II, III	
	Prémios Anf. 11.1.3	**Ciências da Saúde I** Anf. 11.1.10
		Ciências Naturais I Anf. 11.1.12
19:00–21:30	Receção de Boas-vindas	Fábrica Centro Ciência Viva de Aveiro

Sexta-feira, 29 de junho de 2018

9:00–10:20	Sessões Comunicações Orais IV, V e VI		
	Prémios (cont.) Anf. 11.1.3	**Ciências da Saúde II** Anf. 11.1.10	**Ciências Naturais II** Anf. 11.1.12
10:20–11:00	Sessões Convidadas I e II		
	Elizabeth Juarez-Colunga Anf. 11.1.3	***María Xosé Rodríguez-Álvarez*** Anf. 11.1.10	
11:00–11:40	Pausa para Café		Sala de Professores
11:00–11:40	Sessão de Pósteres II		Corredor, piso 3
11:40–12:40	Sessão Plenária II: ***Alessandro Fassò***		Anf. 11.1.3
	Statistical modelling of atmospheric profiles and their uncertainty		
12:40–14:00	Almoço		Cantina do Crasto
14:00–15:20	Mesa-redonda: *O papel do Estatístico nas várias fases do Ensaio Clínico* Participantes: ***João Branco***, ***Elsa Branco*** e ***Aurora Baluja*** Moderador: ***Julio Singer*** Anf. 11.1.3		
15:20–16:00	Sessões Convidadas III e IV		
	Daniel Farewell Anf. 11.1.3	***Ruwanthi Kolamunnage-Donà*** Anf. 11.1.10	
16:00–16:20	Pausa para Café		Bar Matemática
16:20–17:20	Sessão Plenária III: ***Carlos Daniel Paulino***		Anf. 11.1.3
	Água de lastro de navios e sua composição biológica: Verificação do cumprimento de normas internacionais por via bayesiana		
18:15–19:00	Passeio *Ria de Aveiro*		
20:15–22:30	Jantar e Cerimónia de Entrega de Prémios		Meliá Ria Hotel & Spa

Sábado, 30 de junho de 2018

9:20–10:20	Sessões Comunicações Orais VII, VIII e IX		
	Modelos Mistos e Longitudinais Anf. 11.1.3	**Ciências da Saúde III** Anf. 11.1.10	**Estatística Multivariada** Anf. 11.1.12
10:20–11:00	Sessões Convidadas V e VI		
	Luiz Alexandre Peternelli Anf. 11.1.3	***Bruno Falissard*** Anf. 11.1.10	
11:00–11:20	Pausa para Café		Bar Matemática
11:20–12:20	Sessão Plenária IV: ***Peter Müller*** *The future of Bayesian clinical trial design*		Anf. 11.1.3
12:20–12:50	Assembleia Geral Extraordinária da SPE		Anf. 11.1.3
12:50–14:20	Almoço		Cantina do Crasto
14:20–15:00	Sessões Convidadas VII e VIII		
	Lurdes Inoue Anf. 11.1.3	***Raquel Menezes*** Anf. 11.1.10	
15:00–16:00	Sessões Comunicações Orais X e XI		
	Análise de Regressão Anf. 11.1.3	**Análise Espacial** Anf. 11.1.10	
16:00–16:30	Sessão de Encerramento		Anf. 11.1.3

Programa Científico

Quinta-feira, 28 de junho

9:00 – 11:00 e 11:30 – 13:00 **Anf. 11.1.3**

Minicurso Moderadora: *Adelaide Freitas*

BIOMETRY WITH COMPOSITIONAL DATA.

Karel Hron, Universidade de Palacký.

15:00 – 16:00 **Anf. 11.1.3**

Sessão Plenária I Moderadora: *Maria Eduarda Silva*

NONPARAMETRIC INFERENCE AND COVARIATE SIGNIFICANCE TESTS IN MIXTURE CURE MODELS.

Ricardo Cao, Universidade da Corunha.

16:00 – 16:40 **Corredor Sala de Professores, piso 3**

Sessão de Pósteres I Moderadoras: *Amalia Jácome Pumar e Magda Monteiro*

P.1 KIDNEY INSUFICIENCY: A STATISTICAL ANALYSIS BASED ON THE GAMLSS FRAMEWORK.

Ana Julia Righetto, Thiago Gentil Ramires, Luiz Ricardo Nakamura, Edwin M. M. Ortega, Gauss M. Cordeiro.

P.2 INFEÇÕES POR PROTOZOÁRIOS INTESTINAIS E DÉFICE DE CRESCIMENTO EM LACTENTES DE SÃO TOMÉ: UM ESTUDO DE COORTE DE NASCIMENTO.

Marta Alves, Ana Luísa Papoila, Marisol Garzón, Luís Pereira-da-Silva.

P.3 FATORES QUE CONDICIONAM A ACEITAÇÃO DA DIRETIVA DA LINHA DE SAÚDE 24 DE NÃO IR A UM SERVIÇO DE URGÊNCIAS.

Isabel Natário, Paula Simões, Joaquim Pina, Sérgio Gomes.

P.4 SCREENING PROCEDURES BASED IN MODIFIED CLASSIFICATION TREES APPLIED TO PAEDIATRIC FAMILIAL HYPERCHOLESTEROLEMIA.

João Albuquerque, Mafalda Bourbon, Marília Antunes.

P.5 MEDIDAS INFORMATIVAS EM ESTUDOS LONGITUDINAIS: UM ESTUDO DE SIMULAÇÃO.

Adriana Vieira, Inês Sousa.

P.6 ENSAIOS CLÍNICOS: HISTÓRIA E EVOLUÇÃO.

Raquel Correia, Fernanda Diamantino.

P.7 TEMPERATURA À SUPERFÍCIE DO MAR E ÍNDICE DE AFLORAMENTO COSTEIRO: MODELAÇÃO E COMPARACÃO AO LONGO DA COSTA DE PORTUGAL CONTINENTAL.

Bruno Monteiro, M. Rosário Ramos, Clara Cordeiro.

P.8 MODELAÇÃO DE VALORES EXTREMOS DE TENSÃO ARTERIAL.

Constantino Pereira Caetano, Patricia de Zea Bermudez.

P.9 UM CONTRIBUTO DA ANÁLISE ESTATÍSTICA NA GESTÃO DE UMA ESTAÇÃO DE TRATAMENTO DE ÁGUAS RESIDUAIS (ETAR).

A. Manuela Gonçalves, M. Teresa Amorim, Marco Costa.

P.10 MODIFICAÇÃO NO MODELO PROBIT PARA AVALIAÇÃO DA GERMINAÇÃO EM SEMENTES DE MILHO.

Deoclecio Jardim Amorim, Rute Quelvia de Faria, Amanda Rithieli Pereira dos Santos, Maria Márcia Pereira Sartori.

P.11 PREDIÇÃO DINÂMICA DA SOBREVIVÊNCIA A LONGO PRAZO EM DOENTES COM CANCRO DA MAMA.

Sofia Azevedo, Susana Esteves, Lisete Sousa.

P.12 ANÁLISE ESTATÍSTICA DAS TEMPERATURAS MENSAIS DO AR NO PORTO - MODELAÇÃO DE ESPAÇO DE ESTADOS NO PERÍODO DE 1888 A 2001.

Marco Costa, Magda Monteiro.

P.13 MODELOS LONGITUDINAIS PARA MOMENTOS DE INOVAÇÃO EM PSICOTERAPIA.

Gina da Silva Voss, Inês Pereira Silva Cunha de Sousa.

16:40 – 18:00 — Anf. 11.1.3

Sessão de Comunicações Orais I: Prémios — Moderadora: *Inês Sousa*

O.1 UAV FOTOGRAMÉTRICO NA AVALIACIÓN DE MASAS FORESTAIS AFECTADAS POR INCENDIOS.

Laura Alonso, Julia Armesto, Marta Fernández, Juan Picos.

O.2 DETEÇÃO DE GRUPOS DE OBSERVAÇÕES ATÍPICAS: UMA APLICAÇÃO EM DADOS GENÓMICOS.

Ana Tavares, Vera Afreixo, Paula Brito.

O.3 DEPRIVATION-SPECIFIC LIFE TABLES USING MULTIVARIABLE FLEXIBLE MODELLING - TRENDS FROM 2000-2002 TO 2010-2012.

Luís Antunes, Denisa Mendonça, Ana Isabel Ribeiro, Camille Maringe, Bernard Rachet.

O.4 RHYTHMICITY ANALYSIS IN CHRONOBIOLOGY USING ORDER RESTRICTED INFERENCE.

Yolanda Larriba, C. Rueda, M. A. Fernández, S. D. Peddada.

16:40 – 18:00 Anf. 11.1.10

Sessão de Comunicações Orais II: Ciências da Saúde I Moderadora: *M. Salomé Cabral*

O.5 AN APPLICATION OF STRATIFIED BOOTSTRAP IN THE DETERMINATION OF LIPID AND LIPOPROTEIN REFERENCE PERCENTILES FOR THE PORTUGUESE POPULATION.

Cibelle Mariano, Marília Antunes, Mafalda Bourbon.

O.6 HOW ASYMMETRIC IS VOLATILITY IN HRV?

Argentina Leite, Ana Paula Rocha, Maria Eduarda Silva.

O.7 UM MODELO LINEAR MISTO PARA REGRESSÃO SEGMENTADA LINEAR/QUADRÁTICA.

Julio M. Singer, Francisco M.M. Rocha, Antonio Carlos Pedroso-de-Lima, Giovani L. Silva, Giuliana C. Coatti, Mayana Zatz.

O.8 NONPARAMETRIC MIXTURE CURE MODELS WITH CURE PARTIALLY KNOWN.

M. Amália Jácome, I. López-de-Ullibarri.

16:40 – 18:00 Anf. 11.1.12

Sessão de Comunicações Orais III: Ciências Naturais I Moderadora: *Isabel Natário*

O.9 POPULATION DYNAMICS EQUILIBRIUM AND EXTREME GROWTH.

M. Fátima Brilhante, M. Ivette Gomes, Dinis Pestana.

O.10 MODELAGEM DE CAPTURAS EM PESO INFLACIONADAS DE ZEROS NO BAIXO RIO AMAZONAS.

Júlio C. Pereira, Giovani L. Silva, Victória J. Isaac.

O.11 MEDIDAS DE FIABILIDADE DE CLASSIFICAÇÃO BINÁRIA COM BASE NUMA VARIÁVEL QUANTITATIVA - UMA COMPARAÇÃO VIA SIMULAÇÃO.

Rui Santos, Miguel Felgueiras, João Paulo Martins, Liliana Ferreira.

O.12 IMPUTAÇÃO MÚLTIPLA BASEADA NO ALGORITMO MONTE CARLO VIA CADEIA DE MARKOV (MCMC) PARA A ESTIMAÇÃO DE PARÂMETROS GENÉTICOS QUANTITATIVOS E SELEÇÃO DE GENÓTIPOS.

Maria Márcia Pereira Sartori, Lucas Vasconcelos Vieira, Gabriela Nunes da Piedade, Maurício Dutra Zanotto.

Sexta-feira, 29 de junho

9:00 – 10:20 — Anf. 11.1.3

Sessão de Comunicações Orais IV: Prémios (cont.) Moderador: *Javier Roca Pardiñas*

O.13 NONLINEAR BEHAVIOR IN THE CURE FRACTION.

Thiago G. Ramires, Ana Julia Righetto, Luiz Ricardo Nakamura, Rodrigo R. Pescim.

O.14 NOVAS ABORDAGENS PARA MODELAÇÃO DE AMOSTRAGEM PREFERENCIAL NA DIMENSÃO TEMPORAL.

Andreia Monteiro, Raquel Menezes, Maria Eduarda Silva.

O.15 A COMPARATION OF PRESMOOTHING METHODS IN THE ESTIMATION OF TRANSITION PROBABILITIES.

Gustavo Soutinho, Luís Meira-Machado, Pedro Oliveira.

O.16 ON THE PARAMETERS ESTIMATION OF HIV DYNAMIC MODELS.

Diana Rocha, Sónia Gouveia, Carla Pinto, Manuel Scotto, João Nuno Tavares, Emília Valadas, Luís Filipe Caldeira.

9:00 – 10:20 — Anf. 11.1.10

Sessão de Comunicações Orais V: Ciências da Saúde II Moderadora: *Marília Antunes*

O.17 ESTIMATION OF REFERENCE EQUATIONS FOR SPIROMETRY FOR NON-CAUCASION POPULATION.

Carina Silva, Anália Matos, Tânia Duarte.

O.18 APLICAÇÃO DE PATH ANALYSIS NA IDENTIFICAÇÃO DE PREDITORES DA QUALIDADE DE VIDA DE PESSOAS COM DOENÇAS CRÓNICAS.

Estela Vilhena, José Luís Pais Ribeiro, Denisa Mendonça.

O.19 DETERMINAÇÃO DA COMPOSIÇÃO CORPORAL EM JOVENS ADULTOS - AVALIAÇÃO DA REPRODUTIBILIDADE ENTRE PROTOCOLOS ECOGRÁFICOS E IDENTIFICAÇÃO DE PREDITORES DE MASSA GORDA TOTAL.

Mário Monteiro, João Paulo de Figueiredo, Sandra Assunção, Rute Santos, António Figueiredo.

O.20 PERMUTATION DISTRIBUTIONS FOR PATTERN CLASSIFICATION IN NEUROIMAGING.

Mohammed S. Al-Rawi, Adelaide Freitas, João V. Duarte, Miguel Castelo-Branco.

9:00 – 10:20 — Anf. 11.1.12

Sessão de Comunicações Orais VI: Ciências Naturais II Moderadora: *Fátima Brilhante*

O.21 UN MODELO DE OPTIMIZACIÓN CONTINUA MULTIOBJETIVO PARA PLANIFICACIÓN FORESTAL.

José M. González-González, Miguel E. Vázquez-Méndez, Ulises Diéguez-Aranda.

O.22 ESTUDO DE CASO CONTROLO: DILEMAS NO CÁLCULO DO TAMANHO AMOSTRAL.

Luzia Gonçalves.

O.23 AGREEMENT BETWEEN REGIONAL CLIMATE PROJECTIONS FROM DIFFERENT EURO-CORDEX MODELS: AN EXPLORATORY STUDY.

Ana Martins, Sandra Rafael, Alexandra Monteiro, Manuel Scotto, Sónia Gouveia.

O.24 A ESCOLHA ESTATÍSTICA E A ESTIMAÇÃO EM TEORIA DE VALORES EXTREMOS: APLICAÇÃO EM DADOS AMBIENTAIS.

Manuela Neves, Helena Penalva, Sandra Nunes, Dora Prata Gomes.

10:20 – 11:00 — Anf. 11.1.3

Sessão Convidada I Moderador: *Javier Roca Pardiñas*

JOINT MODELING OF LONGITUDINAL AND INTERVAL CENSORED TIME-TO-EVENT OUTCOMES: APPLICATION TO TACROLIMUS AND ANTIBODY FORMATION IN KIDNEY TRANSPLANT PATIENTS.

Elizabeth Juarez-Colunga, Universidade do Colorado.

10:20 – 11:00 — Anf. 11.1.10

Sessão Convidada II Moderadora: *Ana Luisa Papoila*

PENALISED SPLINE ESTIMATION FOR THE TIME-DEPENDENT ROC CURVE IN THE PRESENCE OF EXTERNAL INFORMATION.

María Xosé Rodríguez-Álvarez, Centro Basco de Matemática Aplicada.

11:00 – 11:40 — Corredor Sala de Professores, piso 3

Sessão de Pósteres II Moderadores: *M. José Ginzo Villamayor e Marco Costa*

P.14 IMPLEMENTATION OF BOOTSTRAP METHODS FOR ACCURACY ASSESSMENT OF SPACE-TIME DATA MODELLING.

Gustavo Soutinho, Raquel Menezes.

P.15 ANÁLISE ESPACIAL DAS PARTÍCULAS PM10 NA ÁREA METROPOLITANA DE LISBOA.

Paula Pereira, Conceição Ribeiro.

P.16 MODELOS COMPETITIVOS PARA ANALISAR AS SÉRIES TEMPORAIS DA CONCENTRAÇÃO DO OXIGÉNIO DISSOLVIDO NO RIO VOUGA.

Magda Monteiro, Marco Costa.

P.17 MODELAÇÃO CONJUNTA DE DADOS LONGITUDINAIS E DE SOBREVIVÊNCIA EM DESISTÊNCIA DA PSICOTERAPIA.

Ângela Ferreira, Inês Sousa, Eugénia Ribeiro, Miguel Gonçalves, Paulo Machado.

P.18 NIRS AS A RAPID SCREENING METHOD TO PREDICT FIBER CONTENT IN SUGARCANE.

Mateus Gonçalves, W. J. Cardoso, J. V. Roque, R. A. Ferreira, Luiz Peternelli.

P.19 ANÁLISE DE RISCO NA ATIVIDADE FLORESTAL.

Mónica Rodrigues, Maria da Conceição Costa, Isabel Pereira.

P.20 ANÁLISE DA PRESENÇA DE VARIÁVEIS MEDIADORAS – APLICAÇÃO A DADOS DE UM INQUÉRITO REALIZADO NA CIDADE DA PRAIA EM CABO VERDE.

Catarina Venda, P. de Zea Bermudez, Luzia Gonçalves.

P.21 COMPARAÇÃO BAYESIANA DE TESTES DE DIAGNÓSTICO COM DADOS DENSAMENTE OMISSOS AO ACASO.

Carlos Daniel Paulino, Giovani L. Silva.

P.22 UMA BASE CONCEITUAL RACIONAL PARA O EXPERIMENTO.

João Gilberto Corrêa da Silva.

P.23 TEOR FOLIAR DE NUTRIENTES EM AMENDOIM (*Arachis hypogaea L.*) ASSOCIADOS COM FUNGOS MICORRÍZICOS ARBUSCULARES E SUPLEMENTADOS COM EXTRATO SOLÚVEL DE ALGAS, AVALIADOS POR ANÁLISE MULTIDIMENSIONAL "GLM E CANDISC".

Renata B. S. Coscolin, João R. Favan, Deoclecio Jardim Amorim, Edilson R. Gomes, Fernando Broetto, Maria M. P. Sartori.

P.24 VALIDAÇÃO DE MÉTODOS ECOGRÁFICOS NO ESTUDO DA ARQUITETURA DO MÚSCULO MASSÉTER COM RECURSO A ANÁLISE GRÁFICA DE BLAND-ALTMAN E COEFICIENTE DE CORRELAÇÃO DE CONCORDÂNCIA.

Alexandra André, João de Figueiredo, Luís Camilo, Vanessa Domingues.

P.25 STRUCTURAL EQUATIONS MODEL OF A QUESTIONNAIRE ON THE PATIENT SAFETY CULTURE IN PORTUGUESE PRIMARY CARE.

Carina Silva, Margarida Eiras.

P.26 MODELAÇÃO CONJUNTA DE DADOS LONGITUDINAIS E DADOS DE SOBREVIVÊNCIA NA PRESENÇA DE RISCOS COMPETITIVOS.

Laetitia Teixeira, Inês Sousa, Anabela Rodrigues, Denisa Mendonça.

P.27 FUNÇÃO DE LIGAÇÃO DE CAUCHY PARA AVALIAÇÃO DE P50 DE LONGEVIDADE DE SEMENTES DE SOJA.

Amanda Rithieli Pereira dos Santos, Rute Quelvia de Faria, Deoclecio Jardim Amorim, Edvaldo Aparecido Amaral da Silva,Maria Márcia Pereira Sartori.

11:40 – 12:40 — Anf. 11.1.3

Sessão Plenária II — Moderador: *César Sánchez-Sellero*

STATISTICAL MODELLING OF ATMOSPHERIC PROFILES AND THEIR UNCERTAINTY.

Alessandro Fassò. Universidade de Bérgamo.

14:00 – 15:20 — Anf. 11.1.3

Mesa-redonda — Moderador: *Júlio Singer*

O PAPEL DO ESTATÍSTICO NAS VÁRIAS FASES DO ENSAIO CLÍNICO.

João Branco, Universidade de Lisboa.
Elsa Branco, Novartis - Portugal.
Aurora Baluja, Universidade de Santiago de Compostela.

15:20 – 16:00 — Anf. 11.1.3

Sessão Convidada III — Moderador: *Giovani Silva*

NO SUCH THING AS MISSING DATA.

Daniel Farewell, Universidade de Cardiff.

15:20 – 16:00 — Anf. 11.1.10

Sessão Convidada IV — Moderadora: *Inês Sousa*

EVALUATING THE TIME DEPENDENT EFFICACY OF A LONGITUDINAL BIOMARKER FOR CLINICAL ENDPOINT.

Ruwanthi Kolamunnage-Donà, Universidade de Liverpool.

16:20 – 17:20 — Anf. 11.1.3

Sessão Plenária III — Moderadora: *M. Esther López Vizcaíno*

ÁGUA DE LASTRO DE NAVIOS E SUA COMPOSIÇÃO BIOLÓGICA: VERIFICAÇÃO DO CUMPRIMENTO DE NORMAS INTERNACIONAIS POR VIA BAYESIANA.

Carlos Daniel Paulino, Universidade de Lisboa.

Sábado, 30 de junho

9:20 – 10:20 — Anf. 11.1.3

Sessão de Comunicações Orais VII: Modelos Mistos e Longitudinais

Moderadora: *Laetitia Teixeira*

O.25 EFFECTS OF A HEALTH EDUCATION INTERVENTION ON PHYSICAL ACTIVITY IN INDIVIDUALS WITH MODERATE-TO-HIGH CARDIOVASCULAR RISK.

Lucimere Bohn, Pedro Sa-Couto, Ana Ramoa Castro, Fernando Ribeiro, José Oliveira.

O.26 NONLINEAR MIXED-EFFECTS MODEL FOR CYCLOSPORINE PHARMACOKINETICS IN RENAL TRANSPLANT.

A. Sofia Cardoso, M. Salomé Cabral, A. Paula Carrondo e José Guerra.

O.27 JOINT MODELLING FOR LONGITUDINAL AND TIME-TO-EVENT IN HEALTH SCIENCES: WHERE WE ARE AND POSSIBLE EXTENSIONS.

Inês Sousa.

9:20 – 10:20 — Anf. 11.1.10

Sessão de Comunicações Orais VIII: Ciências da Saúde III — Moderador: *Rui Santos*

O.28 ANALYSIS OF CLUSTERED ORDINAL SPATIAL PERIODONTAL DATA USING A NON-PARAMETRIC SPATIAL MODEL FOR INDEPENDENT LATTICES.

Rui Martins, Vanda Inácio de Carvalho.

O.29 PERCEÇÃO PARENTAL DO PESO E ESTILOS DE VIDA DOS ADOLESCENTES - UMA APLICAÇÃO DE MEDIDAS DE CONCORDÂNCIA ENTRE INQUÉRITOS.

Elsa Silva, Augusta Gama, Marília Antunes.

O.30 AVALIAÇÃO DA ATIVIDADE DO LÚPUS: SLEDAI VS EVA.

Ana Cristina Matos, Carla Henriques, Diogo Jesus.

9:20 – 10:20 — Anf. 11.1.12

Sessão de Comunicações Orais IX: Estatística Multivariada

Moderador: *C. Luis Iglesias Patiño*

O.31 HYPERSPECTRAL IMAGE CLASSIFICATION USING FUNCTIONAL DATA ANALYSIS.

M. Oviedo de la Fuente, M. Febrero-Bande.

O.32 UN TEST ESTADÍSTICO PARA ANALIZAR LA VARIABILIDAD ESPACIAL DE LOS DATOS USANDO COMPONENTES PRINCIPALES GEOGRÁFICAMENTE PONDERADAS.

J. Roca-Pardiñas, C. Ordóñez.

O.33 A INFLUÊNCIA DA GESTÃO NA PRODUTIVIDADE DE PLANTAÇÕES DE *EUCALIPTUS GLOBULUS*.

Catarina Monteiro, Nélia Silva, Isabel Pereira.

10:20 – 11:00 — Anf. 11.1.3

Sessão Convidada V — Moderadora: *Manuela Neves*

POTENCIALIDADE DA ESTATÍSTICA NO MELHORAMENTO DE PLANTAS.

Luiz Alexandre Peternelli. Universidade Federal de Viçosa.

10:20 – 11:00 — Anf. 11.1.10

Sessão Convidada VI — Moderadora: *Lisete de Sousa*

POST-APPROVAL APPRAISAL: WHAT ARE THE MAIN METHODOLOGICAL ISSUES?

Bruno Falissard, Universidade de Paris-Sud.

11:20 – 12:20 — Anf. 11.1.3

Sessão Plenária IV — Moderadora: *Isabel Pereira*

THE FUTURE OF BAYESIAN CLINICAL TRIAL DESIGN.

Peter Müller. Universidade do Texas Austin.

14:20 – 15:00 — Anf. 11.1.3

Sessão Convidada VII — Moderadora: *Denisa Mendonça*

MODELING DISEASE PROGRESSION ON ACTIVE SURVEILLANCE USING A BAYESIAN JOINT LONGITUDINAL COMPETING RISKS SURVIVAL MODEL.

Lurdes Inoue, Universidade de Washington.

14:20 – 15:00 — Anf. 11.1.10

Sessão Convidada VIII Moderadora: *Patricia de Zea Bermudez*

BOOTSTRAP METHODS IN MIXED EFFECTS MODELLING: APPLICATION TO DENGUE FEVER IN THE STATE OF GOIÁS, BRAZIL.
Raquel Menezes, Universidade do Minho.

15:00 – 16:00 — Anf. 11.1.3

Sessão de Comunicações Orais X: Análise de Regressão Moderadora: *Luzia Gonçalves*

O.34 ENTROPIA NORMALIZADA E OUTROS MÉTODOS DE SELEÇÃO DE VARIÁVEIS: UM ESTUDO COMPARATIVO COM DADOS SIMULADOS.

Alberto Oliveira da Silva, Rodney Sousa, Pedro Macedo.

O.35 A LACK-OF-FIT TEST FOR QUANTILE REGRESSION MODELS USING LOGISTIC REGRESSION.

Mercedes Conde-Amboage, Valentin Patilea, César Sánchez-Sellero.

O.36 METODOLOGIAS DE MÁXIMA ENTROPIA NA ANÁLISE DE DADOS EM LARGA ESCALA.

Maria da Conceição Costa, Pedro Macedo.

15:00 – 16:00 — Anf. 11.1.10

Sessão de Comunicações Orais XI: Análise Espacial Moderador: *Rui Martins*

O.37 ANÁLISE EXPLORATORIA DA DISTRIBUCIÓN ESPACIAL DOS CENTENARIOS DA GALIZA.

Carlos L. Iglesias Patiño, M. Esther López Vizcaíno.

O.38 DISTRIBUCIÓN ESPACIO-TEMPORAL DE LA MORTALIDAD POR INFARTO AGUDO DE MIOCARDIO EN GALICIA.

María José Ginzo Villamayor, María Isolina Santiago Pérez, María Esther López Vizcaíno, Rosa Maria Crujeiras Casais.

O.39 INCORPORATING SURVEY WEIGHTS IN SPATIAL MODELLING OF OBESITY AND HYPERTENSION IN SOUTH AFRICA.

Sheyla Cassy, Samuel Manda, Filipe Marques, Maria do Rosário Martins, Pedro Silva.

Resumos

Minicurso

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

BIOMETRY WITH COMPOSITIONAL DATA

Karel Hron[1]

[1]Palacký University, Olomouc, Czech Republic

Compositional data are multivariate observations that carry relative information. They are measured in units like proportions, percentages, mg/l, mg/kg, ppm, and so on, i.e., as data that might obey (or not) a constant sum of components. Due to their specific features, particularly scale invariance and non-negativity, the statistical analysis of compositional data must obey the geometry of the simplex sample space. In order to enable processing of compositional data using standard statistical methods, compositions consisting of D components can be conveniently expressed by means of real vectors of $D-1$ logratio coordinates. Meaningful interpretability of such coordinates is of primary importance in practical applications.
Compositional data occur frequently in Biometry. Possible examples are vegetation structure in different regions, or concentrations of metabolites in human urine. They indicate that we are faced up to compositional data whenever the relative structure of components is of primary interest and the relevant information is contained in ratios between components. In the latter example this view is also supported by presence of *size effect*, related to a different sample volume and/or concentration, where an amount of normalization techniques were developed to remove it from data. Examples are, e.g., the well-known AUC normalization whose aim is to normalize a group of peaks by standardizing the area under the curve (AUC) to the group median, mean or any other proper representation of size. Another approach is represented by expressing values relative to a landmark chemical compound, e.g., normalization of urine metabolites to creatinine. The choice of any such normalization is usually strongly data dependent in practice, which affects the objectivity. The logratio methodology presents a systematic approach for processing of compositional data taking specific nature of these observations into account. Since the compositions are expressed in logratio coordinates, standard multivariate statistical methods can be applied, just by taking care about interpretation of the logratio coordinates. Because most of practical data sets contain outliers, also robust counterparts to these methods can be considered. Another challenge are high-dimensional data (like the mentioned metabolomic data) that require adaptation of specific approaches like partial least squares, or methods for three-way analysis.
Aim of the course is to introduce the logratio methodology of compositional data in the context of Biometry. The first part of the course will be devoted to theoretical aspects of the methodology including principles of compositional data analysis, geometrical representation of compositions, construction of logratio coordinates and their interpretability. In the second part exploratory data analysis including visualization will be presented, followed by concrete statistical methods which are popular also in Biometry, e.g. correlation and regression analysis, principal component analysis, or discriminant analysis, even for high-dimensional data (partial least squares). Also robust counterparts to some of these methods will be discussed. Numerical examples will be presented using the package `robCompositions` of the statistical software R.

PRELIMINARY PROGRAM

1. Concepts and principles of compositional data analysis
2. Software in R: The package `robCompositions`
3. Geometrical properties, logratio coordinates
4. Exploratory data analysis
5. Principal component analysis and compositional biplot
6. Supervised methods (classification)
7. Methods for high-dimensional data

Sessões Plenárias

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

NONPARAMETRIC INFERENCE AND COVARIATE SIGNIFICANCE TESTS IN MIXTURE CURE MODELS

Ricardo Cao[1,2], Ana López-Cheda[1] and M. Amalia Jácome[1]

[1]Research group MODES, CITIC, INIBIC, Department of Mathematics, Universidade da Coruña, Spain
[2]ITMATI

ABSTRACT

A completely nonparametric method for the estimation of mixture cure models is proposed. An incidence estimator is extensively studied and a latency estimator is presented. These estimators, which are based on the Beran estimator of the conditional survival function, are proven to be the local maximum likelihood estimators. Two i.i.d. representations for the incidence and the latency estimators are obtained. Moreover, an asymptotic expression for the mean squared error of the latency estimator is derived, and its asymptotic normality is proven. In addition, bootstrap bandwidth selection methods for each nonparametric estimator are introduced. The proposed nonparametric estimators are compared with existing semiparametric approaches in simulation studies, in which the performance of the bootstrap bandwidth selectors are also assessed. The nonparametric incidence and latency estimators are applied to a dataset of colorectal cancer patients from the University Hospital of A Coruña (CHUAC). Furthermore, a nonparametric covariate significance test for the incidence is proposed. The method is extended to non continuous covariates: binary, discrete and qualitative, and also to contexts with a large number of covariates. The efficiency of the procedure is evaluated in a Monte Carlo simulation study, in which the distribution of the test is approximated by bootstrap. The test is applied to a sarcomas dataset.

Keywords and key sentences: Bandwidth selection, bootstrap, censored data, significance tests, cure models, kernel estimation, survival analysis.

1. MAIN RESULTS

A completely nonparametric method for the estimation of mixture cure models is proposed. The nonparametric estimator of the incidence proposed by Xu and Peng (2014) has been extensively studied by López-Cheda et al (2017a), who also proposed a nonparametric estimator of the latency. These estimators, which are based on the estimator proposed by Beran (1981) for the conditional survival function, are proven to be the local maximum likelihood estimators. An i.i.d. representation is obtained for the nonparametric incidence estimator. As a consequence, an asymptotically optimal bandwidth is found. Moreover, a bootstrap

bandwidth selection method for the nonparametric incidence estimator is proposed. The introduced nonparametric estimators are compared with existing semiparametric approaches in a simulation study, in which the performance of the bootstrap bandwidth selector is also assessed.

The nonparametric latency estimator for mixture cure models has been studied by López-Cheda et al (2017b). An i.i.d. representation was obtained, the asymptotic mean squared error of the latency estimator was found, and its asymptotic normality has been proven. A bootstrap bandwidth selection method was also introduced by these authors and its efficiency was evaluated in a simulation study. The incidence and latency estimators were used to analyze a database of colorectal cancer from the University Hospital of A Coruña (CHUAC).

Covariate significance tests for cure models are limited to parametric and semiparametric methods. Recently López-Cheda et al (2018) have filled this important gap by proposing a nonparametric covariate significance test for the probability of cure in mixture cure models. The procedure is based on the significance test by Delgado and González-Manteiga (2001), and it is extended to non continuous covariates: binary, discrete and qualitative. Its efficiency is evaluated in a Monte Carlo simulation study, in which the distribution of the test is approximated by bootstrap. The method has been applied to a colorectal cancer dataset. The test has been studied when the number of potential covariates is very high via a FDR approach and a applied to a sarcomas dataset with nearly 372,452 metilation covariates.

ACKNOWLEDGMENT

This research has been supported by MINECO grants MTM2014-52876-R and MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

References

[1] Beran, R. (1981). Nonparametric regression with randomly censored survival data (Tech. Rep.). Berkeley: University of California, Berkeley.

[2] Delgado, M. A. and González-Manteiga, W. (2001). Significance testing in nonparametric regression based on the bootstrap. *Annals of Statistics*, 29, 1469-1507.

[3] López-Cheda, A., Cao, R., Jácome, M. A., Van Keilegom, I. (2017a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics & Data Analysis*, 105, 144-165. doi: 10.1016/j.csda.2016.08.002

[4] López-Cheda, A., Jácome, M. A., Cao, R. (2017b). Nonparametric latency estimation for mixture cure models. *Test*, 26 , 353-376. doi: 10.1007/s11749-016-0515-1

[5] López-Cheda, A., Jácome, M. A., Van Keilegom, I., Cao, R. (2018). Nonparametric covariate significance tests for the incidence in cure models. Submitted for possible publication.

[6] Xu, J., Peng, Y. (2014). Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics*, 42, 1–17. doi: 10.1002/cjs.11197.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

STATISTICAL MODELLING OF ATMOSPHERIC PROFILES AND THEIR UNCERTAINTY

Alessandro Fassò[1]

[1]University of Bergamo, Italy

ABSTRACT

Measurement uncertainty of atmospheric profiles obtained by remote sensing and radiosoundings is crucial in climate change studies. This talk discusses some modelling issues related to functional data representation of temperature and humidity profiles, which arise in two applications related to the GAIA-CLIM Horizon 2020 research project.
The first case study is involved in co-location mismatch of two atmospheric observations, typically a satellite profile and a radiosonde profile. The objective is the assessment of the vertical smoothing mismatch uncertainty related to this profile comparison. To see this, radiosondes are harmonised to match the satellite data in a two steps procedure, which is based on a maximum likelihood approach and exploits the measurement uncertainties in a natural way. At the first step, radiosonde profiles are transformed into continuous functions using splines. At the second step, radiosonde profiles are harmonised by considering weighting functions based on the generalised extreme values probability density function with parameters depending on altitude. The variation between harmonised and non-harmonised radiosonde is then informative on vertical smoothing mismatch.
The second case study is related to geographic gaps in radiosonde monitoring networks. In particular, a gap region is defined as an atmospheric region where the spatial prediction uncertainty is high. To do this global bi-daily radiosonde profiles are modelled as a spatio-temporal process with functional values and a functional kriging variance is used to identify the gaps. Techniques for large data sets are considered.

Keywords and key sentences: Functional data, Generalized additive models, Spatio-temporal data, Mixed models.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

ÁGUA DE LASTRO DE NAVIOS E SUA COMPOSIÇÃO BIOLÓGICA: VERIFICAÇÃO DO CUMPRIMENTO DE NORMAS INTERNACIONAIS POR VIA BAYESIANA

Carlos Daniel Paulino[1], Eliardo G. Costa[2] e Julio M. Singer[3]

[1]Centro de Estatística e Aplicações (CEAUL) & IST, Universidade de Lisboa, Portugal
[2]Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Brasil
[3]Departamento de Estatística, Universidade de São Paulo, Brasil

RESUMO

Metodologias para obtenção do tamanho amostral visando estimar a concentração de organismos em água de lastro e verificação do cumprimento de normas internacionais são aqui desenvolvidas sob uma abordagem bayesiana. Descrevem-se os critérios da cobertura média e do comprimento médio de intervalos de credibilidade sob os modelos bayesianos conjugados Poisson/Gama e Binomial negativa/Beta-linha com parâmetro de escala. Considera-se também um estudo de simulação para avaliar os métodos propostos e discutir problemas práticos relacionados com a coleta dos dados.

Palavras-chave: tamanho amostral, critério da cobertura média, critério do comprimento médio, distribuição Poisson, distribuição binomial negativa, intervalo de credibilidade.

1. INTRODUÇÃO

A avaliação de descargas da água de lastro de navios é um assunto de reiterado e crescente interesse porque a possível introdução de espécies invasivas contidas nessa água em outros ecossistemas pode trazer sérias consequências ao nível ambiental, económico e de saúde pública. A norma D-2 da Organização Marítima Internacional (IMO) estipula que a água deslastrada não deve conter por metro cúbico mais do que 10 organismos vivos com dimensão de 50 ou mais micrómetros (50×10^{-3} mm).
Dada a volumosa quantidade de água de lastro transportada por grandes navios, tem-se de recorrer a procedimentos de amostragem para verificar se a referida norma é satisfeita. O processo amostral é baseado em algum modelo estatístico e algum critério segundo o qual se deve calcular o número de alíquotas de água e o respetivo volume necessários para se decidir se a norma D-2 é ou não cumprida. Uma das dificuldades deste processo reside na possível heterogeneidade da concentração de organismos dentro do tanque da água de lastro (Murphy *et al.*, 2002). Para lidar com essa situação, consideraram-se modelos bayesianos assentes numa componente amostral poissoniana para a contagem X de organismos em alíquotas de volume fixado w e em número n a determinar.

O modelo mais simples é o modelo Poisson homogêneo, com concentração média comum por volume, λ, para as várias alíquotas, complementada com uma distribuição *a priori* Gama para esta. O modelo de grau intermédio de dificuldade estende o anterior no sentido de tomar para a componente amostral a distribuição Binomial negativa, que já acomoda uma sobredispersão, acompanhada por uma distribuição *a priori* Beta-linha, com um adicional parâmetro de escala, para a mesma concentração média num volume unitário. O último e mais flexível modelo considera as concentrações médias por volume nas alíquotas como tendo distribuição desconhecida F, sendo-lhe atribuída *a priori* um processo Dirichlet (Costa, 2017). Este modelo assume assim uma natureza semiparamétrica, sendo conhecido como modelo de mistura por processo Dirichlet.

2. DETERMINAÇÃO DO TAMANHO AMOSTRAL

Os critérios que se consideram aqui para a determinação do menor tamanho n da amostra de alíquotas são os de regiões de credibilidade HPD $R(\mathbf{x_n})$ para as concentrações esperadas com cobertura média (ACC) e com amplitude média (ALC) especificadas. Ambos os critérios consistem em encontrar a menor dimensão amostral que assegure que a região de credibilidade HPD para a concentração esperada por alíquota:

- ACC: de tamanho especificado (ℓ) tenha uma probabilidade de cobertura em média para os vários conjuntos de dados possíveis maior ou igual ao grau de credibilidade fixado à partida $(1-\rho)$, *i.e.*

$$\int_{\mathcal{X}^n} \left[\int_{R(\mathbf{x_n})} h(\lambda|\mathbf{x_n})\mathbf{d}\lambda \right] p(\mathbf{x_n})\mathbf{dx_n} \geq \mathbf{1}-\rho;$$

- ALC: de grau de credibilidade especificado $(1-\rho)$ tenha um tamanho C_ρ em média para os vários conjuntos de dados possíveis menor ou igual ao comprimento máximo fixado à partida ($\ell_{\texttt{max}}$), *i.e.*

$$\int_{\mathcal{X}^n} C_\rho\left[R(\mathbf{x_n})\right]\ p(\mathbf{x_n})\mathbf{dx_n} \leq \ell_{\texttt{max}}.$$

Para os modelos paramétricos as regiões de credibilidade HPD para λ são intervalos em que a sua concretização e determinação das correspondentes probabilidades *a posteriori* são feitas por cálculo analítico ou numérico ou via Monte Carlo (Costa *et al.*, 2018). Para o modelo semiparamétrico o cálculo das quantidades necessárias para a implementação dos critérios ACC e ALC é mais intricado. As regiões de credibilidade $R(\mathbf{x_n})$ poderão ser uniões de intervalos disjuntos, sendo então o seu tamanho calculado por soma dos comprimentos dos intervalos componentes. Na fórmula relativa ao critério ACC o integral interior passa a estar ligado ao processo *a posteriori* $F(\cdot|\mathbf{x_n})$, que é uma mistura de processos Dirichlet calculável por Monte Carlo (Costa, 2017).

Uma vez determinados o número $n = n_{min}$ de alíquotas de volume w segundo algum critério e, em sequência, a amostra de contagens $\mathbf{x_n}$, calcula-se a região de credibilidade HPD a $100(1-\rho)\%$ $R(\mathbf{x_n})$. Com base nesta toma-se uma decisão sobre a conformidade ou não do navio inspecionado com a norma D-2 do seguinte modo:

- Se $10 \in R(\mathbf{x_n})$, adia-se a decisão até se obter informação adicional que permita reaplicar o processo;
- Se $10 \notin R(\mathbf{x_n})$ conclui-se por conformidade (desconformidade) se 10 estiver localizado à direita (esquerda) de $R(\mathbf{x_n})$, ou por adiamento se 10 estiver situado entre duas partes disjuntas de $R(\mathbf{x_n})$.

Reproduzem-se na tabela seguinte para fins ilustrativos os resultados do n_{min} obtido pelo critério ACC para o modelo Binomial negativo/Beta-linha definido por

– $X|\lambda,\phi \sim BiN(\phi, \left[1+\frac{w}{\phi}\lambda\right]^{-1})$ em que ϕ (considerado fixado) especifica o grau de sobredispersão - note-se que $E(X|\lambda,\phi) = w\lambda$ e $Var(X|\lambda,\phi) = w\lambda + (w\lambda)^2/\phi$;
– $\lambda|\lambda_0,\theta_0,\phi \sim Be^*(\theta_0, \frac{\theta_0}{\lambda_0}+1; \frac{\phi}{w})$ com o respetivo valor esperado de λ igual a $(\phi/w)\lambda_0$.

Os cenários fixados para $(w,\ell,\phi,\theta_0,\rho)$ estão indicados na tabela e foi tomado $\lambda_0 = 10w/\phi$ para garantir o valor esperado *a priori* da concentração média igual a 10, ou seja, ao limiar estabelecido na norma D-2.

Tabela 1: Tamanho amostral (n) obtido usando o ACC sob o modelo Binomial negativo e distribuição *a priori* Beta-linha com parâmetros de forma ϕ e θ_0, e $\rho = 0,05$.

Volume alíquota (w)	Tamanho intervalo (ℓ)	ϕ	Parâmetro de forma (θ_0)			
			11	25	50	75
0,5	2	1,0	462	457	453	444
		2,5	229	226	222	216
		5,0	152	149	144	140
		7,5	127	124	118	114
		10,0	113	111	106	101
	4	1,0	115	112	106	101
		2,5	57	53	49	43
		5,0	37	34	29	24
		7,5	30	28	22	17
		10,0	27	24	19	14
1,0	2	1,0	426	422	417	414
		2,5	194	191	188	185
		5,0	115	113	111	108
		7,5	90	88	85	82
		1,.0	77	75	72	70
	4	1,0	110	107	102	99
		2,5	49	46	44	41
		5,0	29	27	24	22
		7,5	22	20	18	15
		10,0	19	17	15	12

3. COMENTÁRIOS

Problemas práticos relacionados com a coleta de alíquotas da água de lastro tem sido abordado por vários autores como Gollasch & David (2017), entre outros. Entre esse problemas menciona-se a dificuldade em aceder ao tanque de lastro e a necessidade de submeter as alíquotas amostradas para análise em laboratório. Assim, alguns dos tamanhos amostrais propostos na tabela (*e.g.*, 462) não são factíveis em face da tecnologia atual.
No entanto, pesquisadores do Instituto Oceanográfico da Universidade de São Paulo estão desenvolvendo um sistema na qual parte da água deslastrada será conduzida através de um dispositivo ótico onde os organismos serão contados por um software apropriado. Isto irá permitir a coleta de um número grande de alíquotas ao longo do processo de deslastre. Infelizmente ainda não dispomos de dados experimentais obtidos por meio desse sistema.

Perante a indisponibilidade corrente de dados reais procedeu-se a um estudo de simulação dos modelos paramétricos para avaliação dos métodos propostos que mostrou que as especificações fixadas previamente são satisfeitas pelos tamanhos amostrais obtidos.

AGRADECIMENTOS

Esta pesquisa recebeu apoio das seguintes entidades: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, processos 153526/2014-9 e 3304126/2015-2), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processo 2013/21728-2), Brasil, e também da Fundação para a Ciência e Tecnologia, Portugal (projeto PEst-OE/MAT/UI0006/2014).

Referências

[1] Costa, E.G. (2017). Tamanho amostral para estimar a concentração de organismos em água de lastro: uma abordagem bayesiana. Tese de doutorado. Departamento de Estatística, Universidade de São Paulo, São Paulo.

[2] Costa, E.G., Paulino, C.D. & Singer, J.M. (2018). Verifying ballast water compliance with international standards: a Bayesian approach. Submitted.

[3] Gollasch, S., David, M. (2017). Recommendations for representative ballast water sampling. *Journal of Sea Research* 123, 1–15

[4] Murphy, K.R., Ritz, D. & Hewitt, C.L. (2002). Heterogeneous zooplankton distribution in a ship's ballast tanks. *Journal of Plankton Research* 24, 729-734.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

THE FUTURE OF BAYESIAN CLINICAL TRIAL DESIGN

Peter Müller[1]

[1]Dpt. Mathematics and Dpt. Statistics & Data Science, University of Texas Austin

ABSTRACT

The notion of one treatment serving a large homogeneous patient population is becoming increasingly hard to sustain. Many recent studies are designed to understand and address heterogeneity of patient populations, exploiting features of adaptive treatment allocations, population enrichment and sequential stopping. In an increasing number of studies the discovery of relevant subpopulations for such adaptive treatment is part of the trial design.
We review some novel clinical trial designs that implement such schemes, using examples with increasing levels of adaptation. First, we start the discussion with adaptation based on a patient's first cycle response in a two-cycle treatment. Next we continue with dynamic treatment regimens that include adaptation on the outcome from the initial front-line therapy. The discussion includes an adjustment for lack of randomization in the assignment of later stage salvage therapies. Third, we review a basket trial design for a study of targeted therapies for cancer. In this study adaptation includes the selection of disease, treatment and a patient subpopulation. Common to these examples is the notion of quantifying the value of alternative treatment allocations and outcomes. In all examples we do this using a utility function that formalizes, for example, the tradeoff of toxicity and efficacy outcomes. A last example shows another application of such utility-based designs. This time without the context of adaptation. A common theme in the examples is the use of model-based methods for statistical design and inference, also known as Bayesian methods.

Keywords and key sentences: Dynamic treatment regimen, Utility-based designs, Clinical trial, Bayesian methods.

Sessões Convidadas

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

JOINT MODELING OF LONGITUDINAL AND INTERVAL CENSORED TIME-TO-EVENT OUTCOMES: APPLICATION TO TACROLIMUS AND ANTIBODY FORMATION IN KIDNEY TRANSPLANT PATIENTS

Elizabeth Juarez-Colunga[1]

[1]Dept of Biostatistics and Informatics, Adult and Child Consortium for Health Outcomes Research and Delivery Science (ACCORDS), Children's Hospital Colorado, University of Colorado Anschutz Medical Campus

ABSTRACT

Tacrolimus (TAC) is an immunosuppressant drug given to kidney transplant recipients post-transplant to prevent antibody formation and kidney allograft rejection. The optimal therapeutic dose for TAC is poorly defined and drug therapy requires frequent monitoring of drug trough levels. Analyzing the association between TAC levels over time and the development of potentially harmful de novo donor specific antibodies (dnDSA) is complex because dnDSA is assessed at discrete times. This talk discusses methods for jointly analyzing a longitudinal biomarker (TAC) and an interval-censored time to event outcome (dnDSA). Using data from the University of Colorado Transplant Center, we investigate a shared random effects model with longitudinal and interval censored survival sub-models. Using this model, we develop a dynamic prediction framework to calculate individualized predicted probabilities of dnDSA-free survival for new subjects, based on historical TAC measurements and demographic information.

Keywords and key sentences: Shared random effects, Survival analysis, Longitudinal data, Bayesian analysis, Dynamic prediction.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

PENALISED SPLINE ESTIMATION FOR THE TIME-DEPENDENT ROC CURVE IN THE PRESENCE OF EXTERNAL INFORMATION

María Xosé Rodríguez-Álvarez[1], Thomas Kneib[2] and Vanda Inácio de Carvalho[3]

[1]BCAM - Basque Center for Applied Mathematics & IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
[2]Chair of Statistics, Georg-August-Universität Göttingen, Germany
[3]School of Mathematics, University of Edinburgh, Scotland, United Kingdom

ABSTRACT

This work presents a novel penalised likelihood-based estimator of the cumulative-dynamic time-dependent receiver operating characteristic (ROC) curve. The proposal allows to account for the possible modifying effect of covariates on the accuracy of prognostic biomarkers. We apply our approach to the evaluation of biomarkers for early prognosis of death after discharge in patients who suffered an acute coronary syndrome.

Keywords and key sentences: time-dependent ROC curve; P-splines; hazard function..

1. INTRODUCTION

The ROC curve is the measure of diagnostic accuracy most widely used for continuous biomarkers. However, in many circumstances, the aim of a study may involve prognosis rather than diagnosis. In such cases, the disease status of an individual is not a fixed characteristic but it varies with time (e.g., death and alive). To assess the accuracy of continuous biomarkers for time-dependent disease outcomes, time-dependent extensions of *Sensitivity*, *Specificity* and ROC curve have been proposed (e.g., Pepe et al. 2008). Moreover, it is well known that the accuracy of a biomarker can be affected by external information or covariates, for instance, characteristics of the patient. In these situations, if we failure to incorporate covariate information into the ROC analysis, the marginal or pooled ROC curve could lead to erroneous conclusions, and thus conditional or covariate-specific measures of accuracy are needed. This work focuses on the estimation of the conditional cumulative-dynamic time-dependent ROC curve. In contrast to previous proposals in this setting, our approach (1) allows for non-linear effects of continuous covariates on the accuracy of prognostic biomarkers, and (2) relaxes the proportional hazards assumption.

2. NOTATION AND PRELIMINARIES

Let T denote the time to the event of interest, Y the quantitative biomarker, and $\boldsymbol{X}$ the p-variate vector of covariates we are interested in. The conditional or covariate-specific time-dependent cumulative *Sensitivity* (Se) and dynamic *Specificity* (Sp) are defined as

$$Se^{\mathbb{C}}(\upsilon, t \mid \boldsymbol{x}) = \Pr[Y > \upsilon \mid T \leq t, \boldsymbol{X} = \boldsymbol{x}],$$
$$Sp^{\mathbb{D}}(\upsilon, t \mid \boldsymbol{x}) = \Pr[Y \leq \upsilon \mid T > t, \boldsymbol{X} = \boldsymbol{x}].$$

Thus, the conditional cumulative-dynamic time-dependent ROC curve is

$$\mathrm{ROC}_{t,\boldsymbol{x}}^{\mathbb{C}/\mathbb{D}}(p) = Se^{\mathbb{C}}\left((1 - Sp^{\mathbb{D}})^{-1}(p, t \mid \boldsymbol{x}), t \mid \boldsymbol{x}\right) \text{ with } p \in (0,1).$$

Note that with the cumulative *Sensitivity* and dynamic *Specificity* interest lies in evaluating the discriminatory capacity of the biomarker Y in distinguishing those individuals – with a covariate vector value $\boldsymbol{x}$ – that will experience the event of interest prior to time t (cases) from those with the event after t (controls). Note also that a possibly different ROC curve, and therefore discriminatory capacity, can be obtained for each covariate vector value $\boldsymbol{x}$ and each time point t.
It can be easily shown that the above expressions can be expressed as

$$Se^{\mathbb{C}}(\upsilon, t \mid \boldsymbol{x}) = \frac{\Pr[Y > \upsilon, T \leq t \mid \boldsymbol{X} = \boldsymbol{x}]}{\Pr[T \leq t \mid \boldsymbol{X} = \boldsymbol{x}]} = \frac{\int_{\upsilon}^{\infty}(1 - S(t \mid y, \boldsymbol{x}))\, dF(y \mid \boldsymbol{x})}{\int_{-\infty}^{\infty}(1 - S(t \mid y, \boldsymbol{x}))\, dF(y \mid \boldsymbol{x})}, \tag{1}$$

$$Sp^{\mathbb{D}}(\upsilon, t \mid \boldsymbol{x}) = \frac{\Pr[Y \leq \upsilon, T > t \mid \boldsymbol{X} = \boldsymbol{x}]}{\Pr[T > t \mid \boldsymbol{X} = \boldsymbol{x}]} = \frac{\int_{-\infty}^{\upsilon} S(t \mid y, \boldsymbol{x})\, dF(y \mid \boldsymbol{x})}{\int_{-\infty}^{\infty} S(t \mid y, \boldsymbol{x})\, dF(y \mid \boldsymbol{x})}. \tag{2}$$

where

$$S(t \mid y, \boldsymbol{x}) = P(T > t \mid Y = y, \boldsymbol{X} = \boldsymbol{x}) \text{ and } F(y \mid \boldsymbol{x}) = P(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}).$$

3. PENALISED-BASED ESTIMATOR

Expressions (1) and (2) make it clear that in order to estimate $Se^{\mathbb{C}}$ and $Sp^{\mathbb{D}}$ we simply need an estimator of $S(t \mid y, \boldsymbol{x})$ and $F(y \mid \boldsymbol{x})$. For estimating $S(t \mid y, \boldsymbol{x})$, we assume a regression-type model for the conditional hazard function $\lambda(t \mid y, \boldsymbol{x})$, i.e.,

$$\lambda(t|y, \boldsymbol{x}) = \exp\left(\alpha_0 + h_t(t) + h_y(y) + \sum_{a=1}^{A} f_a(\boldsymbol{x}_a) + f_{y,t}(y,t) + \sum_{b=1}^{B} f_b(t, \boldsymbol{x}_b) + \sum_{c=1}^{C} f_c(y, \boldsymbol{x}_c)\right), \tag{3}$$

where $\boldsymbol{x}_a$, $\boldsymbol{x}_b$ and $\boldsymbol{x}_c$ denote subsets of covariates, and $h_{\{.\}}$ and $f_{\{.\}}$ define generic representations of different types of covariates and effects (linear or parametric, smooth, etc). Note that the inclusion of functions $f_{y,t}$ and f_b allow relaxing the proportional hazards assumption. Estimation is based on the piecewise exponential model (see e.g., Friedman, 1982; Kauermann, 2005). Via a data augmentation strategy, the piecewise exponential approach allows a (penalised) Poisson-maximum likelihood estimation scheme for model (3) in the presence of censored observations. In addition, it also allows using Generalised Linear Array Models (GLAM, Currie et al, 2006) to speed up computation.
In order to estimate $F(y \mid \boldsymbol{x})$, we propose the following model

$$Y \mid (\boldsymbol{X} = \boldsymbol{x}) = \beta_0 + \sum_{v=1}^{V} f_v(\boldsymbol{x}_v) + \varepsilon, \tag{4}$$

with $\boldsymbol{x}_v$ and f_v as defined before. We assume that $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$ and ε is independent of $\boldsymbol{X}$. Thus, $F(y \mid \boldsymbol{x}) = H\left(y - \beta_0 - \sum_{v=1}^{V} f_v(\boldsymbol{x}_v)\right)$, where $H(u) = \Pr(\varepsilon \leq u)$.
For the specification of the (multidimensional) smooth functions involved in models (3) and (4), use is made of penalised splines (P-splines, Eilers and Marx, 1996, 2003), in combination with (tensor-product of) B-spline basis functions. In addition, each smooth function is decomposed into a penalised and an unpenalised component (see. e.g., Currie and Durban, 2002). This decomposition presents several attractive features: (a) redundant components can be easily identified; and (b) generalised linear mixed models estimation techniques can be used. In this work, estimation of models (3) and (4) is done by means of the method described in Rodríguez-Álvarez et al. (2018).

3. APPLICATION

The Global Registry of Acute Coronary Events (GRACE) scoring system is a well-known risk score (biomarker) for early prognosis of death after discharge in patients who suffered from acute coronary syndrome (ACS). In the construction of the GRACE scoring system, the left ventricular ejection fraction (LVEF), a very well-established prognosticator of mortality in the ACS scenario, was not included, mainly due to presence of missing values. Thus, the death risk estimates from the GRACE risk score might be misleading because the LVEF is conspicuous by its absence in the construction of the GRACE. In order to check this hypothesis, we applied the proposal presented in this work with the aim of evaluating the possible effect of the LVEF on the prognostic value of the GRACE risk score.
The study population consists of 3488 consecutive patients admitted due to ACS at the Hospital Clínico de Santiago de Compostela, Spain. The event of interest was all-cause mortality during follow-up (81% censorship). Figure 1 shows the estimated time-dependent area under the ROC curve (AUC) for the GRACE score adjusted by LVEF at $t = 6$, $t = 12$ and $t = 18$ months after discharge. As can be observed, there is a clear effect of the LVEF on the prognostic value of the GRACE risk score. The red lines represent the estimated marginal/pooled time-dependent AUC, i.e., the time-dependent AUC obtained when pooling the data without regard to the LVEF values. These results highlight that not accounting for the possible modifying effect of the LVEF on the prognostic value of the GRACE risk score would yield to optimistic results.

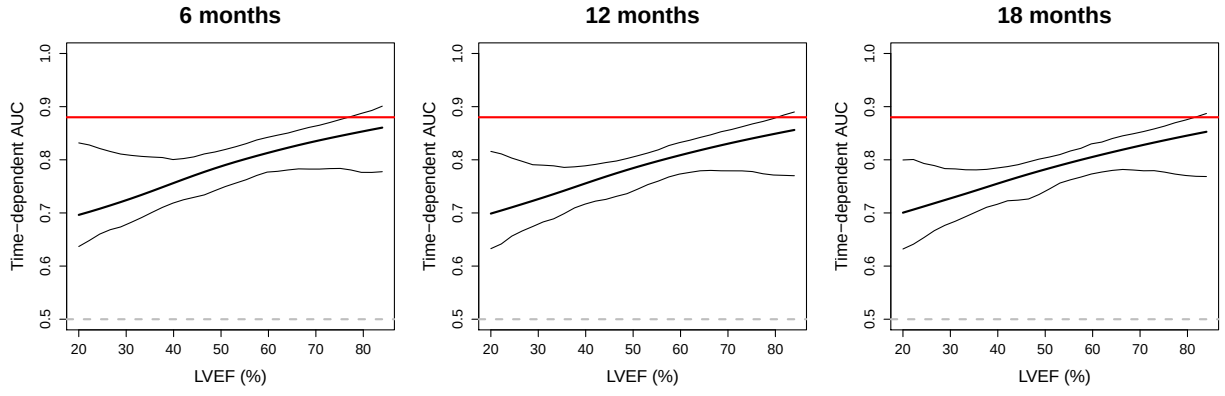


Figure 1: Estimated time-dependent AUCs for the GRACE score adjusted by LVEF(%) at $t = 6$, $t = 12$ and $t = 18$ months after discharge (solid black lines). The red lines represent the marginal/pooled time-dependent AUC.

ACKNOWLEDGEMENT

This research was supported by the Basque Government through the BERC 2018-2021 program and by Spanish Ministry of Economy and Competitiveness MINECO through BCAM Severo Ochoa excellence accreditation SEV-2013-0323 and through project MTM2017-82379-R funded by (AEI/FEDER, UE) and acronym "AFTERAM".

References

[1] Currie, I., and Durban. M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, 4, 333–349.

[2] Currie, I., Durbán, M., and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, 68, 259–280.

[3] Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89–121.

[4] Eilers, P.H.C. and Marx, B.D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, 66, 159–174.

[5] Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10, 101–113.

[6] Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics & Data Analysis*, 49, 169–186.

[7] Pepe, M.S., Zheng, Y., Jin, Y., Huang, Y., Parikh C.R., and Levy, W.C. (2008). Evaluating the ROC performance of markers for future events. *Lifetime Data Analysis*, 14, 86–113.

[8] Rodríguez-Álvarez, M.X., Durban, M., Lee, D.-J., and Eilers, P.H.C. (2018). On the estimation of variance parameters in non-standard generalised linear mixed models: Application to penalised smoothing. *ArXiv: https://arxiv.org/abs/1801.07278*

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

NO SUCH THING AS MISSING DATA

Daniel Farewell[1]

[1]School of Medicine, Cardiff University

ABSTRACT

The phrase "missing data" has come to mean "information we really, really wish we had". But is it actually data, and is it actually missing? I will discuss the practical implications of taking a different philosophical perspective, and demonstrate the use of a simple model for informative observation in longitudinal studies that does not require any notion of missing data.

Keywords and key sentences: Informative drop-out, longitudinal analysis, Missing data.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

EVALUATING THE TIME DEPENDENT EFFICACY OF A LONGITUDINAL BIOMARKER FOR CLINICAL ENDPOINT

Ruwanthi Kolamunnage-Donà[1]

[1]Departament of Biostatistics, Institute of Translational Medicine, University of Liverpool.

ABSTRACT

Many clinical and biomedical studies are aimed at discovering biomarkers that can accurately signal a clinical endpoint. Often such study protocols are based on binary (case/control) disease outcome with a single biomarker measurement at baseline. Although many disease or event-time outcomes and biomarkers are time dependent, complications arise when event-times are censored and biomarkers are measured intermittently. Unless information on longitudinal biomarker and censored event-time processes are combined correctly, the intermittently measured biomarker, measurement error and missingness in biomarker measurement schedule could lead to misleading inference about the regression parameters that describe the true association between a prospective biomarker and subsequent risk of clinical endpoint. In recent years, joint modelling of longitudinal biomarker and event-time processes has gained its popularity as they yield more accurate and precise estimates. Considering this modelling framework, a novel methodology for evaluating the time-dependent efficacy of a longitudinal biomarker for clinical endpoint is proposed and it will assess how well longitudinally repeated measurements of a biomarker over various time periods $(0, t)$ distinguishes between individuals who developed the disease by time t and individuals who remain disease-free beyond time t. The proposed approach is evaluated through simulation and illustrated on the motivating dataset from a prospective observational study of biomarkers to diagnose the onset of sepsis.

Keywords and key sentences: Joint modelling, longitudinal data, event-time data, ROC curve.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

POTENCIALIDADE DA ESTATÍSTICA NO MELHORAMENTO DE PLANTAS

Luiz Alexandre Peternelli[1]

[1]Departamento de Estatística, Universidade Federal de Viçosa, Brasil.

RESUMO

Historicamente a Estatística tem auxiliado pesquisadores a tomar decisões dentro de suas áreas de atuação. Nas Ciências Agrárias o avanço da Estatística na área experimental, principalmente experimentação em campo, permitiu a melhoria de técnicas de cultivo, identificação de produtos mais eficientes para controle de pragas e doenças, adubação do solo para melhor atendimento às necessidades das culturas etc, que contribuíram para o aumento da produtividade desde o início do século passado. Ainda assim haveria necessidade de aumentar ainda mais a produtividade das culturas mais importantes em termos de fornecimento de alimentos. Nesse sentido surgiram teorias e métodos estatísticos associados ao melhoramento genético, tanto vegetal quanto animal, como, por exemplo, os modelos para identificação de genótipos de elevado desempenho em caracteres de interesse agronômico e de indústria, com estabilidade e adaptabilidade adequadas, e a teoria de modelos mistos, que garantiu uma aplicação mais abrangente na área de melhoramento genético. No entanto, outras culturas não tiveram avanços na área de melhoramento baseado no uso de aplicação de metodologias estatísticas de maneira tão efetiva. Em especial citamos o caso da cana de açúcar que até recentemente, em vários programas de melhoramento, tem seus procedimentos de seleção de genótipos baseados na seleção massal, ou seja, na identificação de clones promissores pela simples inspeção visual que os técnicos ditos experientes realizam dentro dos campos de melhoramento. A cana de açúcar é uma das culturas mais importantes na atualidade e na realidade brasileira, não só por ser matéria prima para a fabricação do açúcar e do etanol, mas também por ser uma importante fonte de energia renovável em maior escala. Um programa de melhoramento pode começar com pouco mais de um milhão de indivíduos geneticamente distintos que, ao longo de quase 13 anos, são avaliados, descartados, ou selecionados até o lançamento de uma nova variedade comercial. No entanto, pesquisas que objetivam a seleção precoce de material se faz importante num mundo cada vez mais competitivo e emergencial. A combinação de procedimentos fitotécnicos, estatísticos, genéticos e computacionais tem permitido grande avanço ao encontro desse objetivo. Nesse sentido, cada vez mais informações são coletadas em nível de plantas individuais, nas fases iniciais do programa de melhoramento. Para aumentar a eficiência dos programas de melhoramento, desde o seu início todos os indivíduos podem ser genotipados, visando à obtenção de informações moleculares e suas variantes, ou fenotipados, visando à obtenção de informações

fisiológicas, morfológicas e estruturais em resposta ao meio em que o indivíduo se encontra. Essas informações citadas, quando adicionadas às informações experimentais, ambientais e de pedigree, dentre outras, podem acarretar num imenso volume de dados, estruturados ou não, geralmente complexos, que por sua vez podem impactar na agilidade e eficiência econômica e de resultados em busca de novas variedades. Nessa palestra serão apresentados e discutidos alguns dos problemas que motivam nossa pesquisa, e os métodos estatísticos usados para a sua solução. Todas as pesquisas realizadas visam obter resultados que garantam uma maior eficiência e eficácia do programa de melhoramento da cana desenvolvido na Universidade Federal de Viçosa, Brasil.

Palavras-chave: Modelos mistos, Melhoramento vegetal, Seleção genômica, Componentes de variância, Ciências Agrárias.

AGRADECIMENTOS

CNPq, FAPEMIG, CAPES, RIDESA.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

POST-APPROVAL APPRAISAL: WHAT ARE THE MAIN METHODOLOGICAL ISSUES?

Bruno Falissard[1]

[1]Centre de Recherche en Epidemiologie et Santé des Populations, Université Paris-Sud

ABSTRACT

Reimbursement issues appear more and more challenging as opposed to approval. Curiously, the designs of studies which evaluate drugs efficacy and safety are still those that are required for approval. In addition, the objectives of approval (efficacy and safety, both evaluated in a rather academic way) are really different to those of post-approval (importance of effect, real life effectiveness, economic efficiency). This new context challenges importantly methods that should be used today to evaluate pharmaceutical products. Should new products be evaluated by the firms that have developed them? Is randomization still a necessity? What about sampling? Are statistical models trustworthy in the context of drug appraisal? Who should be able to understand statistical results obtained in evaluation studies? Is Neyman & Pearson lemma still the most appropriate inferential paradigm? The list of critical questions is long and we have to prepare answers already if we want an evolution and not a revolution.

Keywords and key sentences: History of Statistics, Inference, Market access, Randomization, Observational studies.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

MODELING DISEASE PROGRESSION ON ACTIVE SURVEILLANCE USING A BAYESIAN JOINT LONGITUDINAL COMPETING RISKS SURVIVAL MODEL

Lurdes Inoue[1]

[1]Department of Biostatistics, School of Public Health, University of Washington

ABSTRACT

Active surveillance (AS) is increasingly accepted for managing low-risk prostate cancer, but there is no consensus about its optimal implementation due in part to the uncertainty about risks for disease progression. In this talk, we discuss the application of a Bayesian joint model to compare the risks for disease progression from AS studies after accounting for differences in surveillance intervals and competing treatments and evaluate tradeoffs of more versus less frequent biopsies. Our results indicate that men in different AS studies have different risks for biopsy upgrading after variable surveillance protocols and competing treatments are accounted for. Despite these differences, the consequences of more versus less frequent biopsies are similar across AS studies and biennial biopsies may offer an acceptable protocol.

Keywords and key sentences: Joint model, repeated measurements, competing risks survival, active surveillance, Bayesian analysis.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

BOOTSTRAP METHODS IN MIXED EFFECTS MODELLING: APPLICATION TO DENGUE FEVER IN THE STATE OF GOIÁS, BRAZIL

R.Menezes[1], A.Neco-Oliveira[2] and S.Faria[1]

[1]Departamento de Matemática e Aplicações, Universidade do Minho, Portugal
[2]Instituto Federal Goiano, Campus Morrinhos, Brazil

ABSTRACT

Dengue is the most rapidly spreading mosquito-borne viral disease in the world, making it a major public health concern in tropical and subtropical regions. The World Health Organization (WHO) estimates that between 50 and 100 million new infections occur annually in more than 100 endemic countries, including Brazil.

In this work, we aim to study the relation between the number of dengue's notifications in the State of Goiás in Brazil and its climate conditions, taking into account that the latter affect the mosquito's life cycle, virus development, and mosquito-human interactions. The associations between climatic factors and dengue incidence have been explored, but we are unaware of statistical analysis for dengue presence and evolution in this particular region of Brazil. We have available weekly data, collected across 20 cities well-representative of the State Goiás, for the period January 2008 to March 2015.

As our data suggest the presence of spatial and temporal specific effects, we started by investigating alternative generalized linear mixed models to incorporate random effects imposed by a particular *city* or *year*, or even *week*. The covariates explored, taken as fixed effects in our models, include cumulative rainfall, minimum and maximum temperatures, relative humidity and wind speed, at different time-lags. The comparison study also included crossed and nested random effects. Our findings indicate that dengue counts are preferred to be modelled, when considering a nested hierarchical structure with 2 levels, allowing for specific *year*'s effects within each *city*.

At a last stage, assuming a hierarchical 2-levels nested model, we aim to assess the significance of variance components associated to the random effects. We then investigated parametric and non-parametric bootstrap approaches, the latter to avoid distributional assumptions, able to offer accuracy measures for variance estimates of random effects. These are applied to both simulated and real data, allowing us to obtain coverage probabilities above 90%, when the resampling process at the factor's level under study involves up to 50% of the original data. The discussed bootstrap procedures prove to be an useful tool for the analysis of the significance of random effects in the context of generalized linear mixed modelling.

Keywords and key sentences: Generalized Linear Mixed Models; Climate; Dengue; Variance Components; Bootstrap.

1. Motivating example

Dengue fever is a rapidly spread viral disease. All suspected cases must then be notified to the epidemiological surveillance of each municipality by the local health units, via the Notification of Diseases Information System (SINAN). This action allows for the monitoring of the disease transmission patterns and intends to suggest preventive actions to be taken. In Brazil, the first dengue epidemics occurred in 1981-1982 in Boa Vista, and in 1986 in Rio de Janeiro and some Northeastern capitals. Since then, epidemics have occurred associated with the introduction of new serotypes.

The expansion of dengue may be explained by the disorderly growth of urban centers, where more than 80% of the brazilian population is concentrated, and the lack of basic sanitation infrastructures. Furthermore, in many states of Brazil, the climate conditions are deemed favorable to the spread of the disease, making it very difficult to apply public measures capable of eradicating its vector transmitter, the *Aedes aegypti* mosquito, with a worrying increase of dengue cases in young people and children.

The most populous state of center-west region of Brazil is Goiás, with around 6 million inhabitants, distributed among a total of 246 cities, being Goiânia the capital with 1.5 million of people. According to our knowledge, there is a lack of statistical analysis and modelling for dengue presence and evolution in this particular region of Brazil.

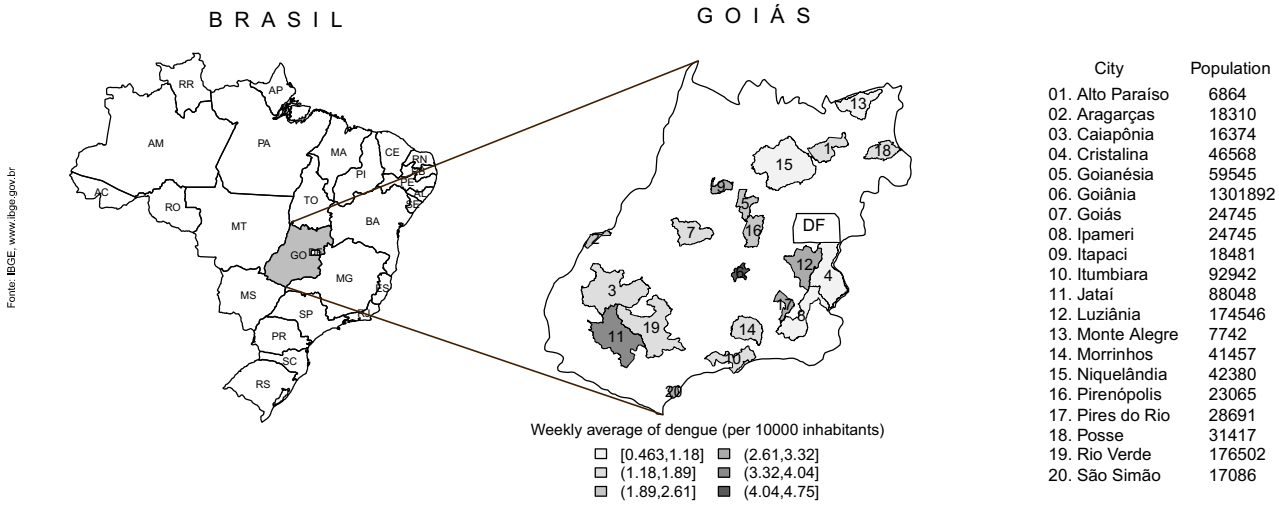


Figure 1: State of Goiás, Brazil. The 20 cities considered in our study.

Table 1: Summary statistics for all variables across the 20 cities, Jan 2008-Mar 2015.

	Variables	**Minimum**	**Mean**	**Maximum**	**Std.Deviation**	**Coef.Variation (%)**
1	*dengue*	0.00	45.38	5145	248.44	547.47
2	*prec (mm)*	0.00	25.06	251.80	36.33	144.97
3	*tmin* (^oC)	1.60	16.70	26.30	3.25	19.46
4	*tmax* (^oC)	23.30	32.35	41.70	2.63	8.13
5	*hram (%)*	18.12	66.15	91.04	14.82	22.40
6	*wind (m/s)*	0.00	1.70	4.77	0.72	42.35

In our study, we considered the number of weekly dengue notifications (dengue) and meteorological data from 20 cities in the state of Goiás (Figure 1), from January 2008 to March

2015 (377 weeks). The meteorological data were obtained from the Brazil National Meteorological Institute, making available information on daily cumulative rainfall (*prec*), minimum and maximum temperatures (*tmin/tmax*), average relative humidity (*hram*) and average wind speed (*wind*) over the period under study. These data were then converted into weekly averages. Summary statistics for these data throughout the entire 377 weeks across all cities are given in Table 1.

2. Comparison study of GLMMs

We first analysed different generalized linear mixed models (GLMMs), to study the relation between the number of dengue's notifications in the state of Goiás and its climate variations. Mixed models are characterized as containing both fixed and random effects. The fixed effects, defined by the meteorological variables at different time-lags, are analogous to standard regression coefficients and are estimated directly. The random effects consider specificities imposed by *city* or *year* (or *week*), and they are summarized in terms of their estimated variances and covariances.
As we wish to model the incidence of dengue as response variable, our comparison study involved the Poisson and the Negative Binomial distributions, specifying an *offset* variable as a function of the population. The second distribution is important to handle data overdispersion, as it happens in our case study, and the corresponding model can be represented as

$$Y_{ijs}|a_i, b_j, c_s \sim NegBin(\mu_{ijs}, k)$$

$$E[Y_{ijs}|a_i, b_j, c_s] = \mu_{ijs}$$

$$Var[Y_{ijs}|a_i, b_j, c_s] = \mu_{ijs} + \frac{\mu_{ijs}^2}{k}$$

$$\eta_{ijs} = \beta_0 + \beta_1 \times prec_{ij(s-lag_1)} + \beta_2 \times tmin_{ij(s-lag_2)} + \beta_3 \times tmax_{ij(s-lag_3)} +$$
$$\beta_4 \times hram_{ij(s-lag_4)} + \beta_5 \times speed_{ij(s-lag_5)} + offset(ln(thab_{ij})) + a_i + b_j + c_s$$

$$ln(\mu_{ijs}) = \eta_{ijs}$$

$$a_i \sim N(0, \sigma_a^2)$$

$$b_j \sim N(0, \sigma_b^2)$$

$$c_s \sim N(0, \sigma_c^2)$$

where Y_{ijs} represents the number of dengue occurrences at *city* $i = 1, \ldots, 20$, *year* $j = 2008, \ldots, 2015$ and *week* $s = 1, \ldots, 52$. The $\eta_{ijs} = log(\mu_{ijs})$ is the linear predictor, with the *offset* variable given by $ln(thab_{ij})$, where $thab_{ij}$ is the total population at *city* i and *year* j. Moreover, c_s is a spatial random effect, a_i and b_j are temporal random effects, *year* and *week* specific, being $c_s \sim N(0, \sigma_c^2)$, $a_i \sim N(0, \sigma_a^2)$ and $b_j \sim N(0, \sigma_b^2)$. The possible time-lags for the meteorological variables are represented by $lag_1, \ldots, lag_5$.
The main results of our comparison study, restricted to crossed random effects, are summarized at Table 2. Based on the Akaike Information Criterium (AIC), model M5 seems to be preferable. So, we next analysed an alternative mixed effects model with a nested hierarchical structure with two levels, allowing for specific *year*'s effects within each *city*, as represented in Figure 2.

Table 2: Results of the modelling comparison study. M0 – M1 are GLMs (restricted to fixed effects) for Poisson and Negative Binomial distributions, respectively. M2 – M6 always assume a response variable following a Negative Binomial distribution, and they consider both fixed and random effects, so they identify GLMMs. The estimates with $*$ are not statistically different from zero for a significance level of 5%.

PARAMETERS	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**
Constant	-13.7679	-12.6564	-12.6231	-12.8740	-11.2647	-11.5195	-10.2061
prec (lag 6)	0.0029	0.0048	0.0037	0.0056	0.0010$*$	0.0045	6.49e-5$*$
tmin (lag 4)	0.1790	0.1422	0.174	0.1277	0.0139$*$	0.1507	0.0531
tmax	0.0337	-0.0261	-0.0378	-0.0129$*$	0.0410	-0.0693	0.0073$*$
hram (lag 10)	0.0340	0.0418	0.0406	0.0395	0.0168	0.0385	0.0067
speed (lag 2)	-0.5174	-0.3183	-0.4725	-0.2940	-0.2073	-0.3309	-0.2778
AIC	303103	38482	37763	37411	38041	36336	37206
k	-	0.4137	0.4837	0.5174	0.4697	0.6550	0.5674
ϕ	42.1643	1.022	1.0152	1.0178	1.0139	1.0044	1.0070
σ_a^2	-	-	0.3358	-	-	0.4111	0.3410
σ_b^2	-	-	-	0.4881	-	0.6598	-
σ_c^2	-	-	-	-	0.6891	-	0.8175

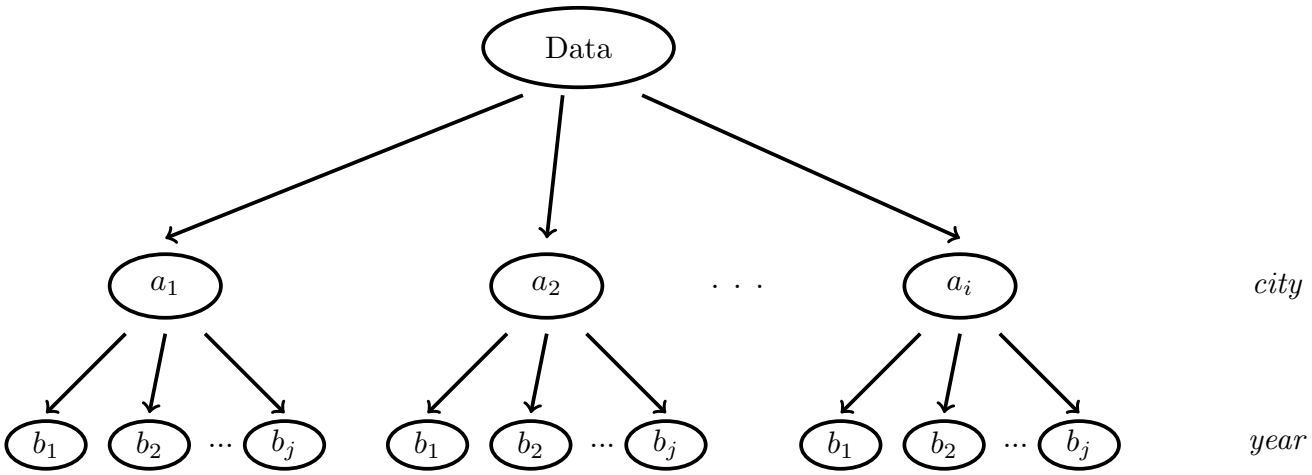


Figure 2: A nested hierarchical structure with two levels.

3. Bootstrap assessment of variance components

Finally, we investigated parametric and non-parametric bootstrap approaches, aiming to assess the significance of variance components associated to the random effects in GLMMs.
We confirmed that the *year* and the location of the *city* are also determining factors in dengue incidence.
Based on the results obtained, the need for public policies, together with joint actions involving local population, are confirmed to be important to combat the dengue fever and avoid epidemic periods.

Mesa Redonda

O PAPEL DO ESTATÍSTICO NAS VÁRIAS FASES DO ENSAIO CLÍNICO

Oradores

- João Branco, Instituto Superior Técnico, Universidade de Lisboa (IST-UL, Portugal)
- Elsa Branco, Country Monitoring Head at Novartis Farma - Produtos Farmacêuticos S.A. (Novartis, Portugal)
- Aurora Baluja, Universidade de Santiago de Compostela (USC, Espanha)

Moderador

- Júlio Singer, Universidade de São Paulo (USP, Brasil)

RESUMO

Antes de um medicamento ser aprovado, precisa passar por várias fases que permitam avaliar a sua eficácia e segurança. A investigação clínica envolve a avaliação de tratamentos médicos propostos, a avaliação dos benefícios de terapias existentes e o estabelecimento de combinações de fármacos e respetivas dosagens.

A Estatística desempenha um papel crucial nos Ensaios Clínicos e no processo de desenvolvimento de medi-camentos - desde o delineamento do ensaio, passando pelo estabelecimento do protocolo, até à análise esta-tística dos resultados obtidos. Atua desde a conceção, condução, análise e relatório do estudo clínico, permi-tindo, por exemplo, o controlo e minimização de vieses e de variáveis de confundimento.

Todos os intervenientes num Ensaio Clínico devem comunicar entre si de modo a garantir um delineamento e uma análise bem-sucedidos. Um fator que muitas vezes dificulta a comunicação efetiva é uma terminologia estatística complicada. Torna-se crucial que a equipa compreenda a estratégia proposta pelo estatístico.

Esta mesa-redonda tem como objetivo debater a importância da Estatística nas várias fases de um Ensaio Clínico, na perspetiva do Estatístico, do Farmacêutico e do Médico. Os principais tópicos a serem abordados incluem a concepção de um ensaio clínico com o consequente cuidado na especificação das hipóteses a serem testadas, sua implementação operacional, a aquisição, armazenamento, confidencialidade e tratamento estatístico dos dados juntamente com o imprescindível rigor em sua documentação. Nesse contexto, problemas associados à inerente multidisciplinaridade serão salientados. Além disso, aspectos relacionados com a formação necessária para aqueles que pretendem trabalhar nessa área serão dicutidos face aos desafios oriundos dos novos métodos de aquisição e tratamento de grandes quantidades de dados.

A discussão terá uma duração aproximada de 80 minutos, sendo 10 min. para a exposição oral de cada um dos 3 intervenientes e o restante tempo para o moderador e para a interação com a plateia.

Sessões Prémios

à melhor Comunicação Oral

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

UAV FOTOGRAMÉTRICO NA AVALIACIÓN DE MASAS FORESTAIS AFECTADAS POR INCENDIOS

Laura Alonso[1], Julia Armesto[1*], Marta Fernández[1] e Juan Picos[1]

[1] Escola de Enxeñería Forestal, Universidad de Vigo, A Xunqueira, CP 36005, Pontevedra
[*] Email: julia@uvigo.es

RESUMO

Debido ao cambio climático e os cambios nos usos do solo os incendios forestais constitúen unha ameaza cada vez maior. É esencial analizar con celeridade as áreas afectadas por incendios forestais para deseñar os correspondentes plans de actuación e recuperación. Este traballo mostra a utilidade dun UAV fotogramétrico de baixo custo no modelado de masas afectadas por incendios forestais. Deseñáronse métodos de análise estatístico das nubes de puntos obtidas que permiten avaliar os cambios na estrutura das masas forestais queimadas.

Palabras e frases chave: Incendios forestais, restauración, UAV, nube de puntos, percentiles.

1. INTRODUCCIÓN

Dende que comezaron os rexistros estímase que en Europa arderon por incendios forestais ao redor de 700.000 hectáreas de bosque [1]. En España, rexistráronse un total de 550.000 incendios forestais dende a década dos 60 hasta 2014, afectando cerca de 7,5 Mha. Segundo fontes oficiais, ao redor do 45% dos incendios teñen lugar en Galicia [2]. Debido ao cambio climático, o risco de incendios esta a aumentar considerablemente, xunto con outros fenómenos como a presenza de pragas [3]. Xustamente o ano 2017 foi unha das temporadas de incendios máis devastadoras. Os eventos máis graves foron os acontecidos no fin de semana do 14 e 15 Outubro de 2017 en Galicia, onde en dous días arderon 49.171 hectáreas [5], rexistráronse 4 falecidos e resultaron afectadas 32 vivendas de primeira ocupación.

No marco actual obsérvase un cambio no escenario posterior aos lumes que precisa unha resposta máis rápida para avaliar as superficies afectadas e facilitar a elaboración de programas de actuación e restauración. Esta necesidade de responder con rapidez require profundizar na investigación de novas técnicas de avaliación de superficies forestais afectadas por incendios. Montealegre *et al*. (2017) realizaron estimacións da severidade post-incendio mediante o emprego de variables derivadas do LiDAR PNOA ("Canopy relief ratio" e "porcentaxe de puntos por riba de 1 metro"), obtendo modelos con precisións do 85,5%. Kane *et al*. (2013) tamén emplearon variables obtidas mediante un sensor LiDAR ("altura de copa" e "fracción de cabida cuberta") para determinar as posibles relacións entre a estrutura espacial dos bosques co grao de severidade lo lume.

No presente traballo se describe o levantamento fotogramétrico realizado mediante UAV de baixo custo dunha parcela incendiada e dunha parcela patrón sen queimar. As imaxes foron procesadas para reconstruír nubes de puntos densas xeorreferenciadas. Se obtiveron unha serie de estatísticos das nubes de puntos na área incendiada e patrón que permitiron detectar as diferenzas entre elas na estrutura vertical da vexetación .

2. CASO DE ESTUDO

A masa incendiada de estudo é unha masa regular de *Eucalyptus globulus* de entre 8 e 10 anos localizada en Fragoselo, Concello de Vigo (España), afectada polo incendio do 15/10/2017, que foi medida o 22/12/2017 (ver figura 1). Previo ao incendio o estrato arbustivo estaba conformado por *Rubus* sp., fentos e materia orgánica morta formando unha masa densa de aproximadamente 1,25 metros de altura. Sobrevoáronse un total de 5,5 hectáreas. Se tomou como masa patrón unha parcela situada en Santa Ana, concello de Pontecaldelas (España), repoboada con *Eucalyptus globulus* sen tratamentos selvícolas, que foi sobrevoada o 07/04/2017. Nesta parcela o estrato arbustivo está conformado por mato de *Ulex europaeus* a distintas alturas e con distribución aleatoria e puntual. O total da superficie analizada foron 4,7 ha.

a.  b. 

Figura 1. Fotografías da parcela afectada por incendio (a) e da parcela patrón (b).

2. MATERIAIS E MÉTODOS

Se realizou un voo da parcela incendiada cun drone Phantom 2 equipado cun sensor RGB GoPro HERO 4 Black. O voo durou 4 minutos 48 segundos. Se tomaron imaxes a 1 fps resultando un total de 288 fotogramas, dos cales se descartaron 113 correspondentes a despegue a aterrizaxe. O procesamento das imaxes realizou co programa Pix4D Mapper®. Mediante técnicas fotogramétricas obtívose unha nube de puntos densa e xeorreferenciada (ver figura 2) onde cada punto medido sobre a superficie dos obxectos da escena ten coordenadas X,Y,Z cartográficas. A partir de esta nube de puntos se obteñen diferentes estatísticos mediante a ferramenta software Lastools©.

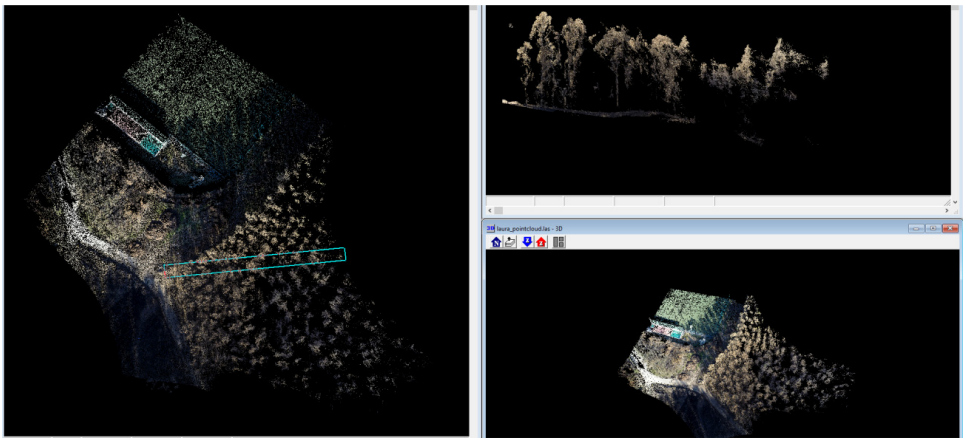


Figura 2: Vista en planta, perfil e 3D da nube de puntos densa obtida co UAV fotogramétrico.

Primeiramente normalizouse a nube de puntos en altura: para cada punto se resta da coordenada Z, a correspondente valor estimada da Z do solo. Esto permite ter alturas comparables. A continuación se obtiveron os estatísticos xerais das nubes de puntos: altura mínima, máxima e media, o cadrado da media, a desviación estándar, a asimetría, a curtosis, os percentís e os bicentís. Os que resultan máis reveladores son a media e os percentís. Esta información permite realizar unha comparación do xeito no que se distribúen os puntos nas dúas masas comparadas, a parcela queimada e a que se utiliza como patrón de comparación. Entre as variables que amosaron gardar unha posible relación co estado das masas, se destacan os percentís, que indican a porcentaxe de puntos que hai por debaixo de certa altura. Mediante esta variable estatística se observou que na masa incendiada, entre o 25% e o 50% dos puntos están por debaixo da altura media (o que correspondería a sotobosque), mentres que na masa patrón esta porcentaxe está desprazada ao 50% e 75% (ver táboa 2).

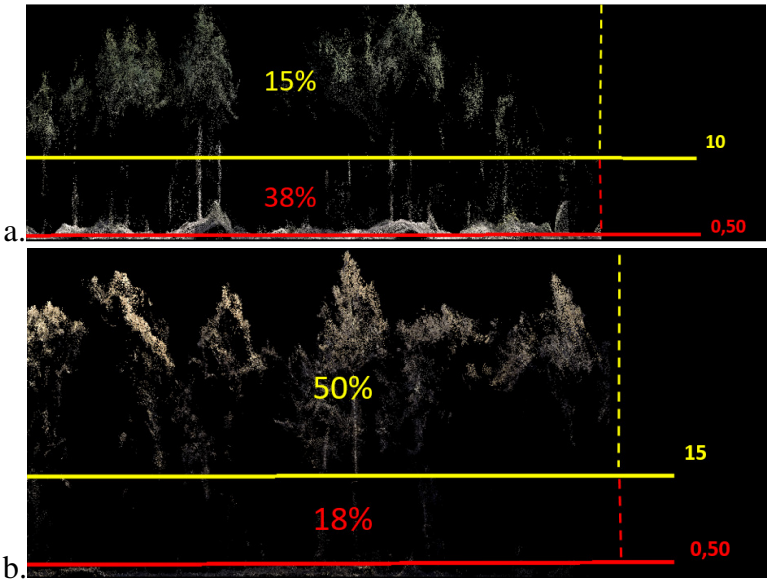
	min	max	avg	qav	std	ske	p05	p10	p25	p50	p75	p90	p95
INCENDIO	1	34	20	431	7	-0,6	6	9	15	21	25	27	29
PATRÓN	1,3	29	8’5	110	6	0,3	1,4	2	2	8	14	17	18

Táboa 2: Estatísticos base obtidos na parcela patrón e na parcela incendio. En cor vermella se encontran sinaladas as alturas medias e percentís entre os que se atopa a altura media en cada parcela.

Esta primeira análise da distribución dos puntos permite inferir que a densidade dos puntos do sotobosque e copas difire entre as masas. Coa fin de confirmar este aspecto se obteñen as porcentaxes de puntos, en ambas parcelas, por riba de diferentes alturas de corte. As alturas de corte escollidas son:

- Altura do solo: 0.5 metros. Se considera que os puntos de altura menor se corresponden a solo.
- Altura de inicio de copa: 15 metros para a parcela incendio; 10 metros para a parcela patrón. Considérase que os puntos comprendidos entre este umbral e a altura do solo corresponden ao sotobosque da masa forestal avaliada.

Obtidas as porcentaxes para estes intervalos de altura observouse que a porcentaxe de puntos en sotobosque difire considerablemente entre a percela patrón e a parcela incendio (ver figura 3). Mentres que na parcela patrón o 38% dos puntos pertence a sotobosque na parcela incendio esta porcentaxe redúcese a un 18%. Con isto podemos obter unha medida da redución do sotobosque tras un incendio.



a.
b.

Figura 3: Porcentaxe de puntos entre altura de solo e o arranque de copa: (a) parcela incendio, (b) parcela patrón.

3. CONCLUSIÓNS

A fotogrametría con UAV permite reconstruír satisfactoriamente nubes de puntos densas de masas forestais abertas e modelar os tres estratos diferenciables: solo, sotobosque e copas. A obtención e selección de estatísticos das nubes de puntos permite obter indicadores do grao de afección do lume na estrutura vertical da masa forestal. No caso de estudio descrito obsérvase que a nube de puntos en parcela queimada se concentra maioritariamente nos estratos copa e solo, mentres que na parcela patrón presenta unha estrutura vertical continua (de copa a solo).

AGRADECEMENTOS

Os autores agradecen ao MINECO e o CDTI (Goberno de España) pola financiación recibida a través do proxecto ITC-20161074, cofinanciado por fondos FEDER.

Referencias

[1] San-Miguel-Ayanz J., Durrant T., Boca R., Libertà G.e Branco A. (2017). Forest Fires in Europe, Middle East and North Africa 2016. *Publications Office*, Luxemburgo.

[2] Barreal J., Loureiro M. e Picos J. (2012). Estudio de la casualidad de los incendios forestales en Galicia. *Economía Agraria y Recursos Naturales,* 99-114.

[3] Rowell A. e Moore P. (2000). Global Review of Forest Fires, WWF; *IUCN*.

[4] Montealegre A. L., Lamelas M. T., Tanase M. A. e De la Riba J. R. (2017). Forest fire severity estimation based on the LiDAR-PNOA data and the values of the Composite Burn Index. *Revista de Teledetección,* nº 49,1-16.

[5] El informe de la Fiscalía sobre la ola de incendios concluye que no hay evidencia de trama (12 de febreiro de 2018). *La Voz de Galicia*. Recuperado de: www.lavozdegalicia.es

[6] Kane V.R., North M.P., Lutz J.A., Churchill D.J, Roberts S. L, Smith D. F., McGaughey R. J., Kane J. T., Brooks M. L. (2014). Assessing fire effects on forest spatial structure using a fusion of Landsat and airborne LiDAR data in Yosemite National Park. *Remote Sensing of Environment,* nº 151, 89-101.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

DETEÇÃO DE GRUPOS DE OBSERVAÇÕES ATÍPICAS: UMA APLICAÇÃO EM DADOS GENÓMICOS

Ana Tavares[1], Vera Afreixo[1] e Paula Brito[2]

[1]Universidade de Aveiro
[2]Universidade do Porto

RESUMO

Este trabalho aborda o problema da deteção de grupos de observações que se afastam da maioria. O objetivo é a deteção de grupos de palavras genómicas cujo padrão de distribuição, ao longo do genoma, se distinga da maioria dos padrões. O método proposto aplica técnicas de classificação hierárquica para identificar classes de pequena dimensão, pois é nesses grupos que se espera encontrar as observações que se demarcam das restantes. Assim, a dimensão das classes serve de ponto de partida para a identificação de potenciais observações atípicas. Num segundo passo, as observações são comparadas com as restantes observações do seu grupo, por forma a avaliar a similaridade entre as distribuições. Para esse efeito é utilizada uma medida de atipicidade funcional que privilegia a forma das distribuições e não apenas a magnitude dos seus valores.

Palavras chave: Deteção de *outliers*, Classificação hierárquica, Distribuições de distâncias, Palavras genómicas.

1. INTRODUÇÃO

O ADN pode ser representado por uma sequência linear composta por quatro símbolos distintos (A, C, G, T). Um segmento de k símbolos consecutivos é designado por palavra genómica. Algumas palavras têm uma função biológica bem definida e muitas regiões funcionalmente importantes do genoma podem ser reconhecidas através da identificação de padrões de sequência, ou "motivos"[5]. Por exemplo, o trinucleótido ATG serve como um local de iniciação nas regiões de codificação (um marcador onde a tradução para a proteína começa) [6]. Conjetura-se que os padrões que têm algum significado funcional ou estrutural estão sujeitos a pressões de seleção positivas (ou negativas) durante a evolução e, consequentemente, têm uma frequência maior (ou menor) do que a esperada [2]. Isso torna a identificação de palavras genómicas e dos seus padrões de distribuição um objeto relevante de investigação.
A análise de sequências de ADN é um domínio de pesquisa amplo e alvo de várias e novas abordagens. Uma dessas abordagens é o estudo das distribuições de distância entre as palavras genómicas [1, 11]. A distância entre palavras (iguais) é definida como a diferença entre a posição dos primeiros símbolos de ocorrências consecutivas da palavra. Por exemplo, as

distâncias entre as palavras $w = AC$ no segmento $ACTGACAGGACAC$ são (4,5,2). A distribuição de distâncias de w traduz a frequência de ocorrência de cada distância, isto é, o número de vezes que a palavra se repete, a uma distância específica, na sequência de ADN em estudo.
Há palavras que revelam, ao longo do genoma, padrões de distribuição que se distinguem da maioria dos restantes, parecendo que foram geradas por um mecanismo diferente. Um exemplo bem documentado na literatura é do dinucleótido CG que, apesar de sub-representado no genoma humano, se aglomera densamente em determinadas regiões (as ilhas de CpG), revelando um perfil de distribuição bem distinto do dos restantes nucleótidos. Por outro lado, existem evidências de que várias palavras genómicas exibem um comportamento específico, que pode ser interpretado como uma assinatura da própria palavra. Palavras com sub-estruturas comuns poderão apresentar padrões de distribuição semelhantes (e.g. ACGCG e CGCGT). Existem motivos que apresentam alguma especificidade funcional, estrutural ou regulatória que são compostos por dezenas de nucleótidos podendo dar origem a um conjunto de palavras de tamanho substancialmente menor que tenham comportamentos semelhantes e bastante específicos. Deste modo, espera-se que, a existir padrões atípicos, estes sejam descritos por um pequeno grupo de palavras e não apenas por uma palavra isolada. Sugere-se que estas palavras são candidatas a apresentar algum tipo de enriquecimento funcional.
Neste trabalho, propomos uma abordagem que recorre a métodos de classificação hierárquica para detetar grupos de pequena dimensão onde se espera encontrar distribuições que se distingam da maioria e sejam, simultaneamente, muito semelhantes entre si.

2. MÉTODO

O método é composto por dois passos, um que agrupa as distribuições de acordo com a sua tendência global; e outro que aplica aos elementos de cada grupo um método de deteção de *outliers* de dados funcionais.
As distribuições empíricas podem apresentar variações que se devem a diferentes causas: variabilidade amostral, mudanças na tendência ou picos de frequência. De modo a evidenciar a tendência global da distribuição é desejável a redução da pequena variabilidade, mantendo a variabilidade mais forte. As distribuições são suavizadas com vista a uma redução da variabilidade amostral, tornando mais evidente a tendência global de cada distribuição. A suavização pode ser considerada uma aproximação à regressão não paramétrica, pelo que não exige pressupostos para a sua utilização. Neste trabalho aplicamos a suavização por funções *spline* cúbicas [9], método bastante flexível na identificação da relação funcional entre as variáveis. Para superar a arbitrariedade da escolha do valor do parâmetro que governa o equilíbrio entre a suavidade da curva e sua proximidade com os valores observados utiliza-se *cross-validation* [4].
A similaridade entre os dados suavizados é explorada através de um método de classificação hierárquica (aglomerativo). As classes de menor dimensão que são agregadas em níveis superiores do dendrograma estão potencialmente associadas aos grupos de interesse. A distância entre as classes formadas em cada etapa do procedimento aglomerativo é avaliada segundo o critério da ligação máxima, ou seja, considera-se a distância entre os dois elementos mais afastados, um de cada classe. A dissimilaridade entre as funções é quantificada usando duas medidas distintas: a distância euclideana e a distância mínima generalizada (Generalized Minimum distance [12]). Esta última quantifica a "massa" que é necessário mover para transformar uma distribuição na outra. Pode dizer-se que a distância euclideana compara distribuições considerando apenas valores correspondentes, enquanto que a distância mínima generalizada considera a dependência entre valores em diferentes pontos do domínio. Os dendrogramas que resultam da classificação hierárquica são cortados levando à obtenção de uma partição. O nível de corte é decidido com base na análise de dois índices de validação interna da partição, o índice de Calinski-Harabasz [3] e o índice de silhueta [7].

Obtida a partição do conjunto de distribuições, explora-se a similaridade entre os elementos de cada grupo, segundo uma abordagem funcional. Assim, uma distribuição pode ser considerada como atípica por dois motivos: por tomar valores que se afastam do intervalo de valores que a maioria das distribuições apresentam (*outlier* de magnitude); ou por exibir uma forma distinta da maioria (*outlier* de forma). Sobre o conjunto de distribuições de cada grupo é aplicado um método de deteção de outliers que compara não apenas a magnitude das distribuições, mas também a sua forma. O método baseia-se na medida *directional outlyingness* [8]. Este passo foca-se essencialmente nos grupos de menor de dimensão, uma vez que é nestes grupos que se espera detetar distribuições atípicas.
O procedimento pode ser repetido em cada um dos grupos de maior dimensão, para averiguar a existência de sub-grupos que, eventualmente, apresentem padrões distintos. A investigação de níveis mais baixos do dendrograma poderá revelar outros pequenos grupos de interesse.
A metodologia é comparada com os resultados obtidos aplicando apenas o referido método de deteção de *outliers* (segundo passo) a todo o conjunto de dados, tal como descrito em [10].

3. APLICAÇÃO A DISTÂNCIAS GENÓMICAS

O procedimento foi aplicado num conjunto de distribuições de distâncias entre palavras. Focamo-nos em palavras de tamanho 5 e nos seus padrões de distribuição ao longo do genoma humano completo. O conjunto é formado por 1024 distribuições e são consideradas distâncias inferiores a 1000.
A análise de classificação hierárquica do conjunto de dados conduziu à obtenção de cinco classes, com 53 (C1), 162 (C2), 705 (C3), 7 (C4) e 97 (C5) elementos, respetivamente. Do ponto de vista gráfico, as distribuições pertencentes às classes C1, C4 e C5 apresentam uma tendência global semelhante dentro de grupo e bem demarcada da tendência dos outros grupos. De facto, nas duas classes de menor dimensão, C1 e C4, não são identificadas distribuições *outliers*; na classer C5 há dois elementos que são marcados como *outliers*, estando muito próximos do limiar entre *outlier* e não *outlier*. Relativamente à classe C2, são marcados como *outliers* 5 observações, duas das quais se afastam claramente dos restantes elementos do grupo. Por fim, é na classe de maior dimensão, C3, que se encontra uma maior variedade de distribuições, sendo 31 marcadas como *outliers*.
Desta primeira análise resulta a existência de dois grupos de distribuições, com padrões semelhantes entre si, mas que se evidenciam dos restantes padrões (C1 e C4). Sendo a classe C3 muito heterogénea, procedemos a uma análise classificatória neste ramo do dendrograma (*subclustering*). A investigação deste nível mais baixo do dendrograma revelou quatro (sub)classes, duas de maior dimensão (sC1: 291, sC2: 403) e duas de dimensão reduzida (sC3: 7, sC4: 4). Nas duas classes de menor dimensão não foram identificadas distribuições outliers.Pelo que, desta segunda análise, resulta a existência de dois grupos de distribuições homogéneas cujos padrões se demarcam dos restantes (sC3 e sC4).

3. CONCLUSÕES

O objetivo do procedimento assenta na determinação automática de grupos de palavras genómicas com padrões de distribuição similares entre si e afastados dos da maioria das palavras. A ideia chave do nosso procedimento consiste em selecionar distribuições pertencentes a classes de menor dimensão e, nestes, avaliar a similaridade entre as distribuições através de um método de deteção e *outliers* funcional.
A aplicação do procedimento no conjunto de palavras de tamanho 5 permitiu identificar quatro grupos de distribuições bem definidos: cada grupo é formado por distribuições com padrões muito semelhantes, uma vez que nenhuma das suas distribuições é marcada como *outlier*. Dois dos grupos identificados são formados unicamente por distribuições marcadas como *outliers* na análise global, o que sugere que o método proposto deteta grupos de distribuições que se destacam da maioria.

AGRADECIMENTOS

Este trabalho é parcialmente financiado pelo FEDER (Fundo Europeu de Desenvolvimento Regional) e FCT (Fundação Portuguesa para a Ciência e Tecnologia) através dos projetos UID/MAT/04106/2013 do CIDMA (Centro de Investigação e Desenvolvimento em Matemática e Aplicações), UID/EEA/50014/2013 do INESC-TEC (Instituto de Engenharia de Sistemas e Computadores do Porto), POCI-01-0145-FEDER-006961 do COMPETE 2020 (Programa Operacional Competitividade e Internacionalização) e bolsa de doutoramento PD/BD/105729/2014 de AT.

Referências

[1] Afreixo, V., Bastos, C. A. C., Pinho, A. J., Garcia, S. P., Ferreira, P.J. S. G.: Genome analysis with inter-nucleotide distances. Bioinformatics. **25** (23), 3064—3070 (2009).

[2] Burge, C., Campbell, A.M., Karlin, S.: Over-and under-representation of short oligonucleotides in DNA sequences. Proceedings of the National Academy of Sciences **89**(4), 1358–1362 (1992)

[3] Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics-theory and Methods **3**(1), 1–27 (1974)

[4] Craven, P., Wahba, G.: Smoothing noisy data with spline functions. Numerische mathematik **31** (4), 377–403 (1978)

[5] MacIsaac, K.D., Fraenkel, E.: Practical strategies for discovering regulatory DNA sequence motifs. PLoS computational biology **2**(4), e36 (2006)

[6] Nakamoto, T.: Evolution and the universality of the mechanism of initiation of protein synthesis. Gene **432**(1), 1–6 (2009)

[7] Rousseeuw, P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)

[8] Rousseeuw, P. J., Raymaekers, J., Hubert, M.: A measure of directional outlyingness with applications to image data and video. Journal of Computational and Graphical Statistics (just-accepted) (2017)

[9] Silverman, B. W.: Some aspects of the spline smoothing approach to non-parametric regression curve fitting. Journal of the Royal Statistical Society. Series B, 1–52 (1985)

[10] Tavares, A. H. M. P., Afreixo, V., Brito, P., Filzmoser, P.: Directional Outlyingness applied to distances between Genomic Words Proceedings 22nd Portuguese Conference on Pattern Recognition. Aveiro. (2016).

[11] Tavares, A. H. M. P., Pinho, A. J., Silva, R. M., Rodrigues, J. M. O. S., Bastos, C. A. C., Ferreira, P. J. S. G., Afreixo, V.: DNA word analysis based on the distribution of the distances between symmetric words. Scientific Report **7** (728), (2017)

[12] Zhao, X., Sandelin, A.: GMD: measuring the distance between histograms with applications on high-throughput sequencing reads. Bioinformatics **28** (8), (2012)

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

DEPRIVATION-SPECIFIC LIFE TABLES USING MULTIVARIABLE FLEXIBLE MODELLING - TRENDS FROM 2000-2002 TO 2010-2012

Luís Antunes[1,2,3], Denisa Mendonça[3,4], Ana Isabel Ribeiro[3], Camille Maringe[5], Bernard Rachet[5]

[1]Department of Epidemiology, Portuguese Oncology Institute – Porto, Portugal
[2]Faculty of Sciences, University of Porto, Portugal
[3]EPIUnit – Institute of Public Health – University of Porto (ISPUP), Porto, Portugal
[4]Institute of Biomedical Sciences Abel Salazar, University of Porto, Portugal
[5]Cancer Survival Group, London School of Hygiene and Tropical Medicine, United Kingdom

ABSTRACT

Mortality data are an important indicator of population health and development. Information on socioeconomic inequalities in mortality is crucial for policy decisions. The aim of this study was to build deprivation-specific life tables using the Portuguese version of the European Deprivation Index (EDI) as a measure of area socioeconomic deprivation, and to evaluate its trends between the periods 2000-2002 and 2010-2012.

Statistics Portugal provided the counts of deaths and population by sex, age group, calendar year and area of residence (parish). A deprivation level was assigned to each parish according to the quintile of their national EDI distribution. Death counts were modelled within the generalised linear model framework, considering a Poisson error with a log link function, using as offset the person-years at risk. Age effect was modelled using restricted cubic splines. Deprivation level, period and interaction between variables were included in the models.

Life expectancy at birth was 74.0 and 80.9 years in $2000 - 2002$, for men and women, respectively, and increased to 77.6 and 83.8 years in 2010-2012. Yet, we observed differences by socioeconomic deprivation: 1.8 and 1.0 years between most and least deprived men and women in 2000-2002. In 2010-2012, the deprivation gap in life expectancy at birth remained similar, at 2.0 and 0.9 years among men and women, respectively. Compared to least deprived, most deprived groups experienced an excess mortality at birth (in 2010-2012, mortality rate ratios of 1.65 and 1.34 in men and women, respectively) which progressively vanished with increasing age.

Substantial and persistent differences in mortality and life expectancy were observed according to area based socioeconomic deprivation. These differences were larger among men and decreased with age for both sexes. No decrease in the deprivation gap was observed between the two periods.

Keywords and key sentences: Life-tables, multivariable modelling, Poisson regression, splines, socioeconomic inequalities in health.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

RHYTHMICITY ANALYSIS IN CHRONOBIOLOGY USING ORDER RESTRICTED INFERENCE

Larriba Y.[1], Rueda C.[1], Fernández M.A.[1] and Peddada S.D.[2]

[1]Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Spain
[2]Department of Biostatistics, School of Public Health, University of Pittsburgh, USA

ABSTRACT

This work presents a novel statistical framework to analyse periodic data in chronobiology. The research has been motivated from problems arising in the analysis of data from phenomena that exhibited temporal rhythmic patterns in oscillatory systems (e.g. circadian clock, cell-cycle). The contributions of the work are twofold. First, a methodology is developed based on a *circular signal* plus error model that is defined using order restrictions. This mathematical formulation of rhythmicity is simple, easily interpretable and very flexible, with the latter property derived from the non-parametric formulation of the signal. Second, using methods based on Order Restricted Inference (ORI), we address various commonly encountered rhythmicity-related problems in practice. Specifically, we develop methodology for detecting rhythmicity in oscillatory systems, especially when times associated with samples are not available. This is a practical problem in a variety of applications, such as when samples are obtained from human biopsies. The proposed methodology is computationally efficient and broadly applicable to address a wide range of questions related to oscillatory systems.

Keywords: Order Restricted Inference, Rhythmicity Detection, Oscillatory Systems.

1. INTRODUCTION

Blood pressure, body temperature or circadian gene-expressions are just a few of the biological phenomena exhibiting rhythmic processes in nature. Such processes display periodic up-down-up patterns, or rhythms, along periods of time, usually of 24 hours length. The study of these temporal rhythms and how they change under different conditions is called chronobiology. For the last two decades, research on chronobiology has had a marked effect on preventing cardiovascular disorders like hypertension, on improving the effectiveness of cancer treatments or on detecting gene-expressions linked to diseases. These and other implications in health motivate a raised interest in chronobiology, in order to identify and/or characterize those rhythmic processes.

From a statistical point of view, the modelling of chronobiological rhythms is a challenge as observed data usually presents low sampling density along a short number of periods. Moreover, rhythm-patterns can adopt a wide range of rhythmic shapes which are not always well

fitted using standard parametric models, as those based on cosine functions, because they are too rigid for rhythms derived from biological systems [1]. In addition, data derived from biological systems are usually related to highly noisy experiments, as in the case of circadian gene-expressions [2].

The main statistical problem to be solved in this context, and the most important one in practice, is that of identifying which patterns on observed data correspond to rhythmic processes and which do not. A wide variety of procedures to detect rhythmicity are available in literature including among them JTK_Cycle (JTK) [3], one of the most popular among biologists, that is based on the Jonckheere-Terpstra test and the Kendall's tau correlation. More recently, [1] presented ORIOS, a novel algorithm based on ORI to detect and classify two-period rhythms derived from the circadian system.

It is important to note that the aforementioned algorithms, as well as most processes in literature to analyse rhythmicity, consider that the timing of the samples (c.f. the time of the day at which samples were taken in the case of the circadian gene-expressions, or the peak time for the genes participating in the cell-cycle) is a priori known. Yet, there exist experiments where the timing of the samples is unrecorded, since it may be impractical or dangerous, as in the case of human organ biopsies. When this happens, the temporal order of samples must be previously estimated before addressing any other question related to rhythmicity. To our knowledge, this relevant rhythmicity question has been barely dealt in literature. Recently, [4] proposed CYCLOPS, a complex approach based on machine learning and neural networks.

In addition to these two major problems (rhythmicity detection and timing estimation), other minor rhythmicity-related problems are raised due to its implications on health. The estimation of the time at which gene-expressions reach their peaks or the rhythm-pattern comparisons along different species are just a few of the additional topics analysed in recent genetics, toxicological or pharmacological studies.

Solutions given until now are specific to the experiment discussed in each paper existing a lack of uniformity in practical procedures. It is highly desirable to establish a general statistical framework to formulate and solve the aforementioned problems that is broadly applicable.

2. CONTRIBUTIONS

The main contribution of this work is the proposal of a mathematical formulation to deal with rhythmicity-related problems using ORI methodology. The key of this formulation is the definition of *circular signal*. Graphically, a *circular signal* can be mapped as a function displaying a temporal up-down-up pattern (see left panel in Figure 1), that underlies in many biological rhythmic processes (see left panels in Figure 2). These patterns, over a discrete number of values, can be defined using order restrictions as follows.

Definition 1. *Circular signal or up-down-up signal*
A signal $\boldsymbol{\mu}$ *in the Euclidean space is said to be circular iff* $\boldsymbol{\mu} \in C = \bigcup_{LU} C_{LU}$, *where* $L, U \in \{1, \ldots, n\}$, $C_{LU} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_1 \leq \cdots \leq \mu_U \geq \cdots \geq \mu_L \leq \cdots \leq \mu_n \leq \mu_1\}$ *if* $L > U$ *and* $C_{LU} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_1 \geq \cdots \geq \mu_L \leq \cdots \leq \mu_U \geq \cdots \geq \mu_n \geq \mu_1\}$ *if* $L < U$.

To analyse observed data, we will consider the (*circular*) *signal* plus error statistical model. The appropiate statistical procedure to make inferences on those models formulated using restrictions, as the model proposed here, is Isotonic Regression (IR) [5]. IR is defined as the solution of a least squared minimization problem that looks for the best fit to a model incorporating restrictions among the parameters. Hence, IR provides an estimator for *circular signal*. To our knowledge, the IR problem for *circular signals* has not been studied in

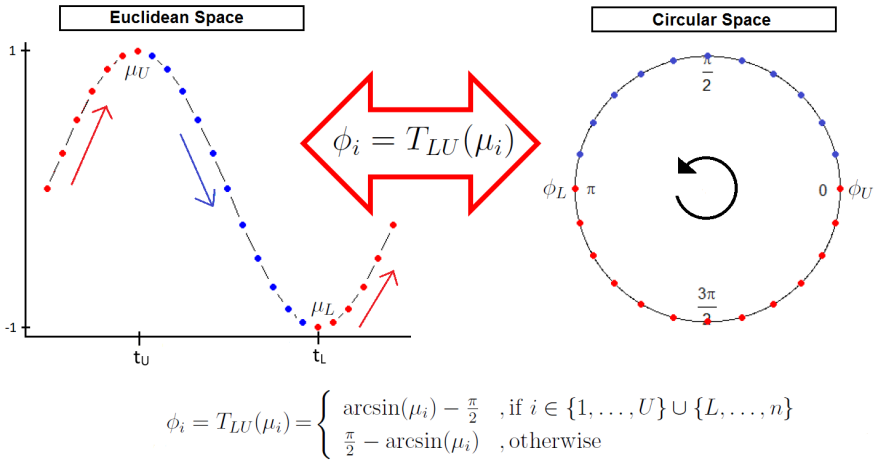


Figure 1: Equivalent formulation of circular signal.

literature. Thus, another contribution of this work is the development and implementation, on the statistical software R, of a computationally efficient algorithm that computes IR for *circular signals*.

The IR *circular signal* estimator is the key to solve many rhytmicity-related problems such as rhythmiticy detection, which is formulated as an hypothesis testing problem contrasting a plain signal pattern against a *circular signal*. Our proposal to solve this test, that incorporates restrictions on the alternative hypothesis, is to conduct a conditional test [6], based on the maximum likelihood ratio test. Minor rhythmicity-related problems, such as peak time estimation or rhythm-pattern comparisons along different species, are likewise solved using the proposed ORI methodology. Specifically, these problems, are tackled as pointwise estimation and as mean comparison testing problems, respectively.

Other interesting feature, thoroughly analysed in this work, that characterises *circular signals* is that they can be equivalently formulated within the Circular space (see Figure 1), from which its name is derived. The rigorous use of the circular geometry is fundamental to formulate and solve problems such as timing estimation when, due to experimental conditions, sample timing is unrecorded. The novel formulation allows us to define the model within the Circular space and to formulate the timing estimation problem as the problem of deriving the optimal circular order among the observed data. The latter problem can be solved as a minimization problem that is itself approximated by related Travelling Salesman Problem.

3. RESULTS

Our new methodological proposals are validated and compared against popular alternatives in literature, which differ from one problem to the other, using simulated and real data. Due to space limitations, only two those analysis have been included here. To assess the performance of ORI methodology detecting rhythmicity, an artificial database is designed imitating what occurs in practice. It contains 15000 simulated patterns, of which 20% are rhythmic ones. The left group of patterns in Table 1 generates rhythmic genes and the right one non rhythmic ones. From Table 1, ORI outperforms JTK on detecting, among others, genes that display asymmetric but rhythmic patterns while controlling misclassification rates. Figure 2 displays the real (left) expression patterns of the genes *Per2* (up) and *Per3* (down) from NIH3T3 mouse liver cell lines (NCBI GEO accession number GSE11922) together with reordered patterns according to ORI (middle) and CYCLOPS (right) timing estimates, under

the assumption that the moment at which samples were taken is unknown. From Figure 2, ORI temporal order estimates provide clearly rhythmic patterns closer to the real ones than CYCLOPS does.

Table 1: False Negative (FNR) and Discovery (FDR) Rates at nominal level of $\alpha = 0.01$.

FNR (for Rhythmic Patterns)						FDR (for Non Rhythmic Patterns)	
Cosine		*Cosine Two*		*Asymmetric*		*Flat*	
ORI	JTK	ORI	JTK	ORI	JTK	ORI	JTK
0.000	0.000	0.000	0.000	0.000	0.956	0.025	0.000

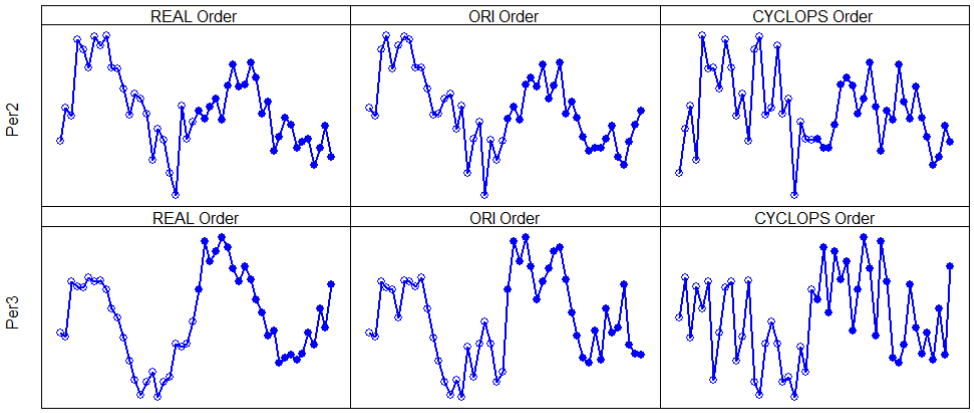


Figure 2: *Per2* and *Per3* gene-expressions plotted using three different orders.

ORI methodology not only provides more efficient solutions to the rhythmicity-related problems described in this work, but also, promising expectations can be glimpsed on it. Due to its flexibility and versatility, we expect the ORI methodology to provide good solutions to many other chronobiological problems such as, for example, the analysis of rhythmic data in the Circular space including covariates in the model, which will be part of our future work.

References

[1] Larriba Y., Rueda C., Fernández M.A., Peddada S.D. (2016). Order restricted inference for oscillatory systems for detecting rhythmic signals. *Nucleic Acids Research* 44, e163.

[2] Larriba Y., Rueda C., Fernández M.A., Peddada S.D. (2018). A bootstrap based measure robust to the choice of normalization methods for detecting rhythmic features in high dimensional data. *Frontiers in Genetics* 9, 24.

[3] Hughes M.E., Hogenesch J.B., Kornacker K. (2010). JTK CYCLE: An Efficient Nonparametric Algorithm for Detecting Rhythmic Components in Genome-Scale Data Sets. *Journal of Biological Rhythms* 25, 372–380.

[4] Anafi R.C., Francey L.J., Hogenesch J.B., Kim J. (2017). CYCLOPS reveals human transcriptional rhythms in health and disease. *Proceedings of the National Academy of Sciences of the United States of America* 20, 5312–5317.

[5] Robertson T., Wright F.T., Dykstra R.L. (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, New York.

[6] Menéndez J.A., Salvador B. (1991). Anomalies of the likelihood ratio tests for testing restricted hypothesis. *The Annals of Statistics* 19, 889–898.

III Encuentro Luso-Galaico de Biometría
Aveiro, 28 - 30 de Junio de 2018

NONLINEAR BEHAVIOR IN THE CURE FRACTION

Thiago G. Ramires[1], Ana Julia Righetto[2], Luiz Ricardo Nakamura[3] and Rodrigo R. Pescim[4]

[1]Universidade Tecnológica Federal do Paraná - Apucarana-PR, Brazil
[2]Instituto Agronômico do Paraná, Londrina-PR, Brazil
[3]Universidade Federal de Santa Catarina, Florianópolis-SC, Brazil
[4]Universidade Estadual de Londrina, Londrina-PR, Brazil

ABSTRACT

Nonlinear effects between explanatory and response variables are increasingly present in new surveys. Here, we propose a flexible three-parameter semi-parametric cure rate survival based on the Weibull distribution. The proposed model is based on the generalized additive models for location, scale and shape, for which any or all parameters of the distribution are parametric linear and/or nonparametric smooth functions of explanatory variables. The new model is used to fit the nonlinear behavior between explanatory variables and cure rate. The flexibility of the proposed model is illustrated by predicting lifetime and cure rate proportion as well as identifying factors associated to women diagnosed with breast cancer.

Keywords and key sentences: Survival analysis; Cure rate models; GAMLSS; P-spline; Residual analysis

1. INTRODUCTION

The objective of this study is to analyze censored data with the presence of cure fraction in which explanatory variables have nonlinear effects in relation to the failure times. Regression models with cure fraction are characterized by a significant fraction of individuals that do not experience the event of interest, even after a long follow-up period. In many cases, explanatory variables can present indefinite behavior. A natural question that arises is how to deal with nonlinearity in the relationship between the outcome variable and a continuous predictor. The incorrect assumption of linearity can lead to a misspecified final model in which a relevant/irrelevant variable may not be included/excluded due to the fact that the hypothesis tests of the parameters related to such variables are based on the slope of the estimated line.
One possible solution would be use categorization, in which such predictors are entered into stepwise selection procedures as linear terms or as dummy variables obtained after grouping. The main issue in the categorization method is that it introduces problems of defining cut-points [1], over-parametrization and loss of efficiency [3, 4]. As an alternative, non-parametric nonlinear functions, such as splines, can be used to estimate different behaviors. Although these methods are relatively advanced, usually such techniques are only adopted on location

or location-scale models, thus requiring the expansion of such techniques to other kinds of models like long-term survival.

2. METHODOLOGY

Considering the mixture models, the survival function for the Weibull cure rate (Weibullcr) model is defined as

$$S_{pop}(t;\mu,\sigma,\nu) = \nu + (1-\nu)\exp(-(t/\mu)^{\sigma}), \tag{1}$$

where $\mu > 0$ and $\sigma > 0$ are the scale and shape parameters, respectively, and ν is the cure rate parameter. The probability density function (pdf) corresponding to (1) is given by

$$f_{pop}(t) = (1-\nu)\frac{\sigma}{\mu}\left(\frac{t}{\mu}\right)^{\sigma-1}\exp\left[-\left(\frac{t}{\mu}\right)^{\sigma}\right]. \tag{2}$$

Let $T \sim Weibullcr(y;\theta)$, where $\theta = (\mu,\sigma,\nu)^T$ denotes the vector of parameters of the pdf (2). Consider independent observations t_i conditional on the parameter vector $\boldsymbol{\theta}_i$ (for $i = 1,\ldots,n$) having pdf $f(t_i;\boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}^T = (\boldsymbol{\mu}^T,\boldsymbol{\sigma}^T,\boldsymbol{\nu}^T)$ is a vector of parameters related to the response variable. The generalized additive models for location, scale and shape [5] (GAMLSS) allow the user to model all parameters in $\boldsymbol{\theta}$ as linear, nonlinear parametric, nonparametric (smooth) function of the explanatory variables and/or random effects terms. We can define semi-parametric structures for the elements of the vector $\boldsymbol{\theta}$ using appropriate link functions as

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\sigma} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} g_1\left(\mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1})\right) \\ g_2\left(\mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} h_{j2}(\mathbf{x}_{j2})\right) \\ g_3\left(\mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} h_{j3}(\mathbf{x}_{j3})\right) \end{bmatrix}, \tag{3}$$

where $g_k(\cdot)$ for $k = 1,2,3$ denote the link functions, $\boldsymbol{\beta}_k = (\beta_{0k},\beta_{1k},\ldots,\beta_{m_k k})^T$ is a parameter vector of length (m_k+1), m_k denotes the number of explanatory variables related to the kth parameter and $\mathbf{X}_k$ is a known model matrix of order $n \times (m_k+1)$. Here, $h_{jk}(\mathbf{x}_{jk})$ are smooth functions of the explanatory variables $\mathbf{x}_{jk}$ for $j = 1,\ldots,J_k$. Here, we consider the penalized splines [2] as smooth functions $h_{jk}(\cdot)$.

4. RESULTS AND DISCUSSIONS

In this section, we predict disease-free survival time (death, second malignancy or cancer recurrence considered as event) by means of a data set corresponding to women diagnosed with breast cancer in Germany [7]. The data comprises 686 node positive women who had complete data for these predictors. These women experienced 299 (43.6%) events during a median follow-up time of 53.9 months, leaving all other patients with a right censored failure time. The explanatory variables measures in the study are described as: $\boldsymbol{t_i}$: recurrence free survival time (in days); $\boldsymbol{\delta_i}$: failure indicator (0: censored, 1: observed); ***age***: age (in years); ***htreat***: hormonal treatment with tamoxifen (0: no, 1: yes); ***menostat***: menopausal status (1: premenopausal, 2: postmenopausal); ***tumsize***: tumor size (in mm); ***tumgrad***: tumor grade, a ordered factor at levels $(1 < 2 < 3)$; ***posnodal***: number of positive lymph nodes; ***prm***: progesterone receptor (in fmol); ***esm***: estrogen receptor (in fmol).
Using the StepGAIC procedure (for further details, please check [6]) to select the additive terms for the different parameters, we provide results for the semi-parametric Weibullcr model. The model parameters are defined by

$$\mu_i = \exp[\beta_{01} + \beta_{11}tumgrad_2 + \beta_{21}tumgrad_3 + pb(esm) + pb(age)],$$
$$\sigma_i = \exp[\beta_{02} + \beta_{12}tumgrad_2 + \beta_{22}tumgrad_3],$$
$$\nu_i = \text{logistic}[\beta_{03} + \beta_{13}prm + \beta_{23}tumsize + \beta_{33}htreat_1 + \beta_{43}tumgrad_2 + \beta_{53}tumgrad_3 + pb(age)],$$

where logistic$(x) = \exp(x)/[1+\exp(x)]$ and pb defines a penalized spline. We also present the results of the parametric Weibullcr model, where $pb(x)$ is replaced by $\beta\, x$. Table 1 provides the maximum likelihood estimates (MLEs), standard errors (SEs) and p-values obtained from the fitted parametric and semi-parametric Weibullcr GAMLSS. The coefficients of the smoothing terms have been omitted (to avoid misinterpretations). We conclude that the explanatory variables *age*, *esm* and *tumgrad* are significant to fit the location parameter, only *tumgrad* is significant to explain the variability on t_i and *prm*, *tumsize*, *htreat*, *tumgrad* and *age* are significant to fit the cure rate parameter, where *age* has a nonlinear effect on it. Also, we may note in this table that *esm* has no significant effect in the usual method (parametric model). In addition, with the global deviance (GD) and AIC statistics, we can conclude that the semi-parametric model provides a better fit.

Tabla 1: MLEs of the parameters, approximate SEs and p-values from the fitted parametric and semi-parametric Weibullcr GAMLSS.

semi-parametric				parametric			
Parameter	Estimate	SE	p-value	Parameter	Estimate	SE	p-value
β_{01}	7.003	0.126	<0.001	β_{01}	6.632	0.124	<0.001
β_{11}	-0.353	0.002	<0.001	β_{11}	-0.122	0.050	0.015
β_{21}	-0.487	0.001	<0.001	β_{21}	-0.455	0.067	<0.001
$pb(esm)$	$df = 6.921$			β_{31}	0.000	0.001	0.7679
$pb(age)$	$df = 6.708$			β_{41}	0.011	0.002	<0.001
β_{02}	1.218	0.046	<0.001	β_{02}	1.061	0.098	<0.001
β_{12}	-0.383	0.046	<0.001	β_{12}	-0.443	0.089	<0.001
β_{22}	-0.383	0.046	<0.001	β_{22}	-0.560	0.101	<0.001
β_{03}	1.394	0.688	<0.001	β_{03}	1.947	0.528	0.002
β_{13}	0.002	0.001	<0.001	β_{13}	0.003	0.001	<0.001
β_{23}	-0.037	0.010	<0.001	β_{23}	-0.031	0.008	<0.001
β_{33}	0.678	0.208	0.012	β_{33}	0.600	0.181	0.001
β_{43}	-0.276	0.212	<0.001	β_{43}	-1.010	0.204	<0.001
β_{53}	-0.070	0.351	<0.001	β_{53}	-0.684	0.255	<0.007
$pb(age)$	$df = 4.204$			β_{63}	-0.021	0.008	0.013
AIC=5181.1		GD=5125.4		AIC=5207.2		GD=5175.2	

In Figure 1, we present the estimated cured proportions, for the semi-parametric model, for different levels of the explanatory variables as function of *age*. We may conclude that the probability of cure is higher for patients age around 45 years, the probability of cure increases when age increases in the range [20,45], then start to decreases. Figure 2 also shows the estimated cured proportions, but now considering the parametric model.

4. CONCLUSION

The semi-parametric *Weibull cure rate* regression model provides a flexible regression model for a dependent real outcome. A real data set is used to illustrate the usefulness of this model, showing that it provides better performance than the usual methods in the presence of nonlinear effects in the cure rate proportion.

References

[1] Altman D.G., Lausen B., Sauerbrei W., Schumacher M. (1994). Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, 86, 829–835.

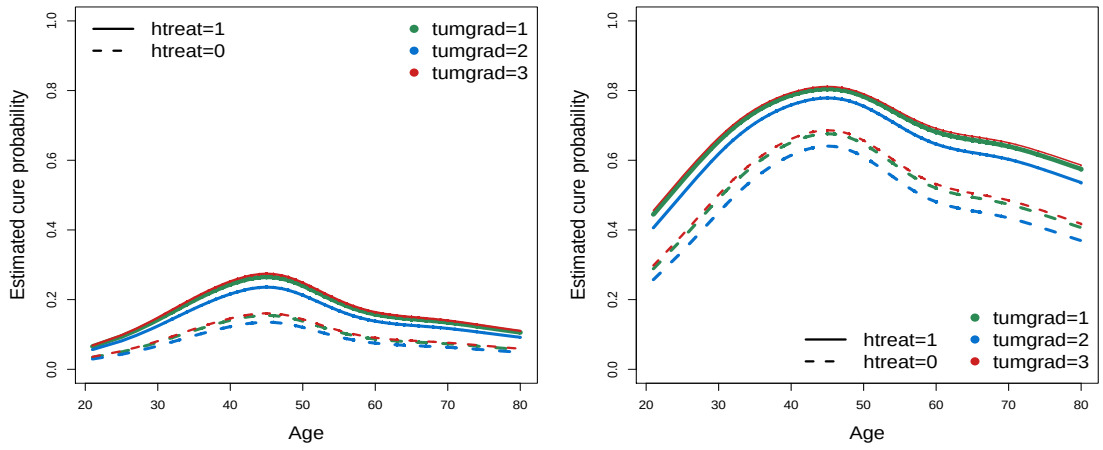


Figura 1: For the semi-parametric model, the estimated cured proportions for each level of *tumgrad* and *htreat* as function of *age* by taking: (a) $\min(prm) = 0$ and $tumsize = 60$ and (b) $prm = 200$ and $tumsize = 10$

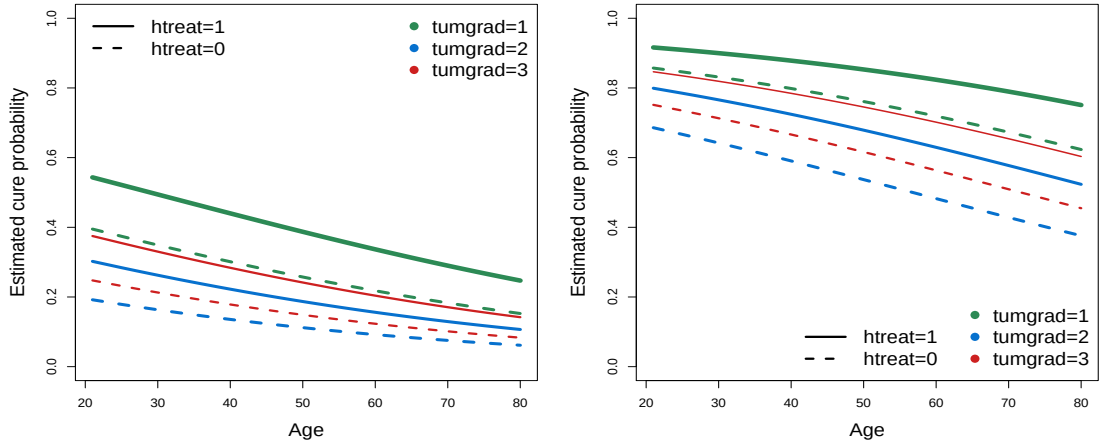


Figura 2: For the parametric model, the estimated cured proportions for each level of *tumgrad* and *htreat* as function of *age* by taking: (a) $\min(prm) = 0$ and $tumsize = 60$ and (b) $prm = 200$ and $tumsize = 10$

[2] Eilers P.H., Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science,* 11, 89–121.

[3] Lagakos S.W. (1988). Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in Medicine*, 7, 257–274.

[4] Morgan T.M., Elashoff, R.M. (1986). Effect of categorizing a continuous covariate on the comparison of survival time. *Journal of the American Statistical Association*, 81, 917–921.

[5] Rigby R.A., Stasinopoulos D.M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 54, 507–554.

[6] Stasinopoulos D.M., Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software,* 23, 1–46.

[7] Schumacher M., Bastert G., Bojar H., Huebner K., *et al.* (1994). Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. *Journal of Clinical Oncology*, 12, 2086–2093.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

NOVAS ABORDAGENS PARA MODELAÇÃO DE AMOSTRAGEM PREFERENCIAL NA DIMENSÃO TEMPORAL

Andreia Monteiro[1], Raquel Menezes[2] e Maria Eduarda Silva[3]

[1]Universidade do Minho & CIDMA
[2]Universidade do Minho
[3]Faculdade de Economia da Universidade do Porto & CIDMA

RESUMO

A análise de dados experimentais que foram observados em diferentes pontos no tempo leva a problemas específicos na modelação estatística e na inferência. Tradicionalmente, nas séries temporais, o foco principal é no caso em que uma variável contínua é medida em pontos temporais discretos equiespaçados, [4]. Existe uma extensa literatura na análise de séries temporais igualmente espaçadas, ver por exemplo [1]. No entanto, dados amostrados de forma irregular ocorrem naturalmente em muitos domínios científicos. Desastres naturais tais como terremotos, inundações ou erupções vulcânicas ocorrem tipicamente em intervalos de tempo irregulares.

Existem poucos métodos para a análise de séries irregularmente espaçadas. Um deles é tornar o esquema de amostragem como regularmente espaçado, usando alguma forma de interpolação. Esta abordagem tem sido aplicada com sucesso a dados irregularmente espaçados causados por valores em falta. Enquanto que esta metodologia, parece ser razoável para lidar com as pequenas irregularidades nos tempos de amostragem causadas por valores em falta, o procedimento de interpolação normalmente altera a dinâmica do processo subjacente, levando a estimativas enviesadas para os parâmetros. Além disso, há pouco consenso na escolha do método específico de interpolação e de qual é o mais apropriado para um determinado conjunto de dados. Como alternativa, pode ser assumido para o processo subjacente um modelo em tempo contínuo, por exemplo um modelo CARMA. Em [4] faz-se uma revisão da aplicação das técnicas do filtro de Kalman para a estimação de processos CARMA.

Um caso particular de dados recolhidos de forma irregular é o de dados em que a sua recolha ao longo do tempo, depende, por determinadas razões, dos próprios valores observados. Por exemplo, um determinado indicador médico do estado de saúde de um indivíduo pode ser medido em diferentes intervalos de tempo e com diferentes frequências, dependendo do próprio estado de saúde. Num contexto completamente diferente, os tempos de ocorrências das transações nos mercados financeiros dependem em larga medida do valor dos ativos subjacentes. Desta forma, informação adicional do fenómeno em estudo é obtida a partir da frequência ou dos tempos de ocorrência das observações. Nestas situações, há uma

dependência estocástica entre o processo que vai ser modelado e os tempos das observações.

Este problema foi primeiramente identificado no contexto da estatística espacial, por [2], que lhe atribuiu o nome de amostragem preferencial. Diggle e os seus coautores demonstraram que ignorar a natureza preferencial da amostragem pode levar a estimativas enviesadas. O nosso trabalho estende o conceito de amostragem preferencial à componente temporal. Relativamente ao modelo proposto, na nossa primeira abordagem, os parâmetros eram estimados por máxima verossimilhança e com recurso a simulações de Monte Carlo. Pretendemos agora apresentar novas propostas, com base em métodos numéricos modernos, como alternativa ao método de Monte Carlo utilizado inicialmente.

Palavras chave: Amostragem Preferencial; Séries Temporais; Processos Autoregressivos em Tempo Contínuo; SPDE.

1. MODELO

Em séries temporais, os dados são obtidos por amostragem de um fenômeno $S(t) : t > 0$ num conjunto discreto de tempos t_i, $i = 1, \ldots, n$. Admitindo a possibilidade de que o desenho amostral possa ser estocástico, $T = (t_1, \ldots, t_n)$ denota o processo estocástico dos tempos de observação. Em muitas situações, $S(t)$ não pode ser medido sem erro, portanto, se Y_i denota o valor medido no tempo t_i, um modelo para os dados assume a forma:

$$Y(t) = \mu + S(t) + N(0, \tau^2),\ t > 0 \tag{1}$$

onde μ é um efeito médio constante e $S(\cdot)$ é o processo Gaussiano estacionário com $E[S(t)] = 0$. Uma formulação equivalente é a de que condicionadas a $S(\cdot)$, Y_i são variáveis mutuamente independentes com distribuição normal com média $\mu + S(t_i)$ e variância τ^2.
Consideramos que:

- S é um processo autoregressivo em tempo contínuo de ordem 1, um $CAR(1)$ que satisfaz a equação diferencial $dS(t) + \alpha_0 S(t)dt = dW(t)$ onde, α_0 é uma constante e $W(t)$ é o movimento Browniano com variância σ_w^2;
- $Y = (Y_1, ...Y_n)^t \sim MVN\left(\mu_y \mathbf{1}, \Sigma_y\right)$

  com $\mu_y = \mu\mathbf{1}$ e $\Sigma_y = \frac{\sigma_w^2}{2\alpha_0} R_y(\alpha_0) + \tau^2 I_n$ em que $\mathbf{1}$ é um vetor de 1's, I_n é a matriz identidade $n \times n$ e $R_y(\alpha_0)$ é uma matriz $n \times n$ em que o elemento $(i,j)^{th}$ é $\rho(|t_i - t_j|)$ definido por $\rho(h) = \frac{\gamma(h)}{\gamma(0)} = e^{-\alpha_0|h|}$

Admitindo que os tempos de amostragem são estocásticos, um modelo completo exige a especificação da distribuição conjunta de S, T e Y. Considerando a dependência estocástica entre S e T, o modelo para lidar com amostragem preferencial é definido através de $[S, T, Y]$ escrita como $[S][T\,|S][Y\,|S(T)\,]$, onde [.] significa a "distribuição de"e escreve-se $S = \{S(t) : t > 0\}$, $T = (t_1, ..., t_n)$, e $S(T)$ representa $\{S(t_1), ..., S(t_n)\}$.
Definimos desta forma uma classe específica de modelos através das suposições adicionais: T é um processo de Cox log-Gaussiano e, T condicionado a S é um processo de Poisson não homogéneo com intensidade $\lambda\,(t) = \exp\{\alpha + \beta S\,(t)\}$. Condicionado a S e T, Y é um conjunto de variáveis gaussianas mutuamente independentes, $[Y_i|S(t_i)] \sim N\left\{\mu + S(t_i), \tau^2\right\}$, com τ^2 igual à variância do erro de medição.
A função de verossimilhança observada para T e Y pode ser escrita como

$$L(\boldsymbol{\theta}) = [T, Y] = \int_S [S][T, Y|S]dS = \int_S [S][T|S][Y|T, S]dS = \int_S [S][T|S][Y|T, S]dS \quad (2)$$

Na nossa primeira abordagem a otimização de (2) era feita com recurso a uma aproximação de Monte Carlo. Neste trabalho, pretendemos apresentar novas propostas, nomeadamente:

- Trabalhar diretamente com (2) evitando a anterior aproximação de Monte Carlo;
- Recorrer ao método proposto em [3], baseado na teoria de equações diferenciais parciais estocásticas (SPDE), para aproximar o processo S definido em tempo contínuo, a um novo processo de Markov definido em tempo discreto. Tal é feito através da compatibilidade integrada do Template Model Builder (TMB) com o R-INLA, que permite criar uma *mesh* temporal e as correspondentes componentes da matriz de precisão esparsa de um campo aleatório Markoviano Gaussiano na dimensão temporal;
- Melhorar significativamente a otimização da função de verossimilhança recorrendo à linguagem de programação C++.

Espera-se que essas mudanças resultem num grande aumento na estabilidade das estimativas dos parâmetros, bem como numa maior eficiência computacional, particularmente em comparação com o nosso método anterior baseado na aproximação de Monte Carlo, bem como com o método do filtro de Kalman adotado pelo pacote cts do R.

Referências

[1] Brockwell, P.J., Davis, R.A.(2016). *Introduction to time series and forecasting.* Springer.

[2] Diggle, P.J., Menezes, R., Su, T.l.(2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*; 59(2), 191–232.

[3] Lindgren F., Rue, H., Lindström, J. (2011). *An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach.* Journal of the Royal Statistical Society: Series B (Statistical Methodology);73(4):423–498.

[4] Tómasson, H.(2015). Some computational aspects of gaussian carma modelling. *Statistics and Computing* ;25(2):375–387.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

A COMPARATION OF PRESMOOTHING METHODS IN THE ESTIMATION OF TRANSITION PROBABILITIES

Gustavo Soutinho[1], Luís Meira-Machado[2] and Pedro Oliveira[3]

[1]University of Minho, Portugal.
[2]Centre of Molecular and Environmental Biology & Department of Mathematics and Applications, University of Minho, Portugal.
[3]EPIUnit, ICBADS, University of Porto, Portugal

ABSTRACT

One major goal in clinical applications of multi-state models is the estimation of transition probabilities. In a recent paper, landmark estimators were proposed to estimate these quantities, and their superiority with respect to the competing estimators has been proved in situations in which the Markov condition is violated. The idea behind their estimator is to use a procedure based on (differences between) Kaplan-Meier estimators derived from a subset of the data consisting of all subjects observed to be in the given state at the given time. Because of this, the computation of their estimator is performed in small sample sizes providing large standard errors in some circumstances. A valid approach is to consider a modification of the landmark estimator based on presmoothing. In this two presmoothing methods are compared. Simulation results indicate that both methods may be much more efficient than the unsmoothed estimator. Real data illustration is included.

Keywords and key sentences: Kaplan-Meier, Multi-state model, Nonparametric estimation, Presmoothing, Survival Analysis.

1. INTRODUCTION

Multi-state models can be successfully used to model the movement of patients among a set of several states. The so-called 'illness-death' model plays a central role in the theory and practice of these models. In these models one important goal is the estimation of the transition probabilities since they allow for long-term predictions of the process. These quantities have been traditionally estimated by the Aalen-Johansen estimator (Aalen and Johansen, 1978), which is consistent if the process is Markovian. Alternative landmark estimators which are consistent regardless the Markov conditions have been proposed in the recent literature (de Uña-Álvarez and Meira-Machado, 2015), and their superiority with respect to the competing estimators has been proved in situations in which the Markov condition is violated. The idea behind the proposed methods is to use specific subsamples or portions of data at hand (namely, those observed to be in a given state at a pre-specified time point) for which the

ordinary Kaplan-Meier (Kaplan and Meier, 1958) survival function leads to a consistent estimator of the target. A weakness of the new method emerges, in some circumstances, from the large estimated standard errors. To avoid this problem, a valid approach is to consider a modification of the landmark estimator based on presmoothing (Cao et al., 2005).

2. PRESMOOTHED ESTIMATORS OF THE TRANSITION PROBABILITIES

A multi-state model is a model for a time continuous stochastic process $(Y(t), t \in \mathcal{T})$ which at any time occupies one of a few possible states. In this paper we consider the progressive illness-death model depicted in Figure **??** and we assume that all subjects are in State 1 at time $t = 0$ (i.e., $Y(0) = 1$). This model is encountered in many medical studies for describing the progression of patients undergoing a given illness, particularly in cancer studies.

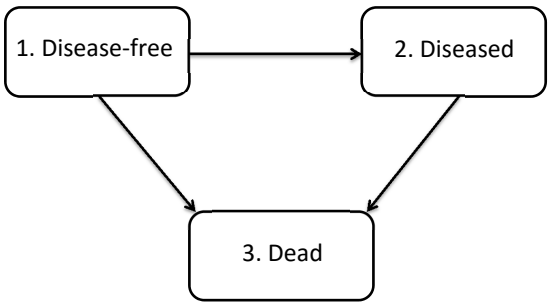


Figure 1: The progressive illness-death model.

For two states k, j and two time points $s < t$, introduce the so-called transition probabilities $p_{kj}(s,t) = P(Y(t) = j|Y(s) = k)$. In the illness-death model we have five different transition probabilities to estimate: $p_{11}(s,t)$, $p_{12}(s,t)$, $p_{13}(s,t)$, $p_{22}(s,t)$ and $p_{23}(s,t)$.
Recently, de Uña-Álvarez and Meira-Machado (2015) introduce new (landmark) estimators for non-Markov processes. The idea of the new methods is to use a procedure based on (differences between) Kaplan-Meier estimators derived from a subset of the data consisting of all subjects observed to be in the given state at the given time. In this paper, we use presmoothing to improve efficiency of the landmark estimators.
Several successful applications of presmoothed estimators have been used in recent literature. All these references concluded that the presmoothed estimators have improved variance when compared to purely nonparametric estimators. This 'presmoothing' is obtained by replacing the censoring indicator variables in the expression of the Kaplan-Meier weights by a smooth fit. This preliminary smoothing may be based on a certain parametric family such as the logistic, probit or cauchit, or on a nonparametric estimator of the binary regression curve. When the parametric family is the right one, parametric presmoothing leads to more efficient estimation than that associated to the unsmoothed estimator. Nonparametric presmoothing is useful when there is a clear risk of a miss-specification of the parametric model. The validity of a given parametric model for presmoothing can be checked graphically or formally, by applying a goodness-of-fit test such as the test proposed by Hosmer and Lemeshow (1989).

3. EXAMPLE OF APPLICATION

In this section we use data of 929 patients affected by colon cancer that underwent a curative surgery for colorectal cancer. In this study, 468 developed recurrence and among these 414 died. 38 patients died without recurrence. Recurrence can be considered as an intermediate transient state and modeled using an illness-death model with transient states 'alive and disease-free' and 'alive with recurrence', and an absorbing state 'dead'.

Figure **??** reports estimated transition probabilities for $p_{ij}(s,t)$, for a fixed value for $s = 365$ days along time t. Plots labbeled as 'Unsmoothed' correspond to the original unsmoothed landmark estimator proposed by de Uña-Álvarez and Meira-Machado (2015) which reveals higher variability on the right hand side. Remaining curves correspond to estimators with a preliminary presmoothing based on a parametric binomial family ('logit', 'probit' or 'cauchit'), on an additive logistic model ('logit.gam') or on a nonparametric regression model ('nonparametric') using the Nadaraya-Watson kernel estimator. We have applied the goodness-of-fit test proposed by Hosmer and Lemeshow (1989) which revealed that the test was able to reject the logistic model when used to presmoothed estimation of the transition probabilities $p_{1j}(365,t)$. Note that the choice of this parametric model is a common choice for a parametric presmoothing. Though one could consider a different parametric model, nonparametric presmoothing is a useful approach when there is a clear risk of a miss-specification of the parametric model.

Plots for the transition probabilities $p_{11}(365,t)$ and $p_{22}(365,t)$ reveal minor differences in the estimated curves based on different methods. Some differences are observed in the right tail of the curves obtained from the different methods when estimating the transition probabilities $p_{12}(365,t)$ and $p_{13}(365,t)$. Plots on bottom of the left-hand side allow for an inspection along time of the probability of being alive with recurrence for the individuals who are disease-free one year, one year after surgery. Since the recurrence state is transient, this curve is first increasing and then decreasing. Major differences are observed in the right tail when comparing the methods based on a parametric presmoothing with their counterparts. A similar behavior is observed when estimating the transition probability $p_{13}(365,t)$ reported in the right-hand side of Figure **??** (top). These plots report one minus the survival fraction along time, among the individuals in the recurrence state. Curves depicted in Figure **??** reveal that the nonparametric presmoothed landmark estimator provide in all cases reliable curves, similar to those obtained from the nonparametric estimators but with less variability. Since there is a clear risk of a miss-specification of the parametric model, the use of estimators based on nonparametric presmoothing as those labelled with 'nonparametric' are preferable.

3. CONCLUSIONS

There have been several recent contributions for the estimation of the transition probabilities in the context of non-Markov multi-state models. Recently, the problem of estimating the transition probabilities in a non-Markov illness-death model has been reviewed, and new estimators have been proposed which are built by considering specific subsets of individuals (namely, those observed to be in a given state at a prespecified time point s for which the ordinary Kaplan-Meier survival function leads to a consistent estimator of the target. As a weakness, it provides large standard errors for large values of s and higher censoring percentages. In this article we compare several approaches based on a presmoothed version of the Kaplan-Meier estimator that can be used to reduce the variability of the proposed estimator. Results obtained in simulations studies (not reported here) suggest that presmoothed approaches are preferable to the original nonparametric estimator, since they often have less variance while providing more reliable curves. Parametric presmoothing is a recommended approach if there is no clear risk of miss-specification of the parametric model. Otherwise the use of nonparametric pressmoothing or based on an additive model is recommended.

ACKNOWLEDGMENT

This research was financed by Portuguese Funds through FCT - "Fundação para a Ciência e a Tecnologia".

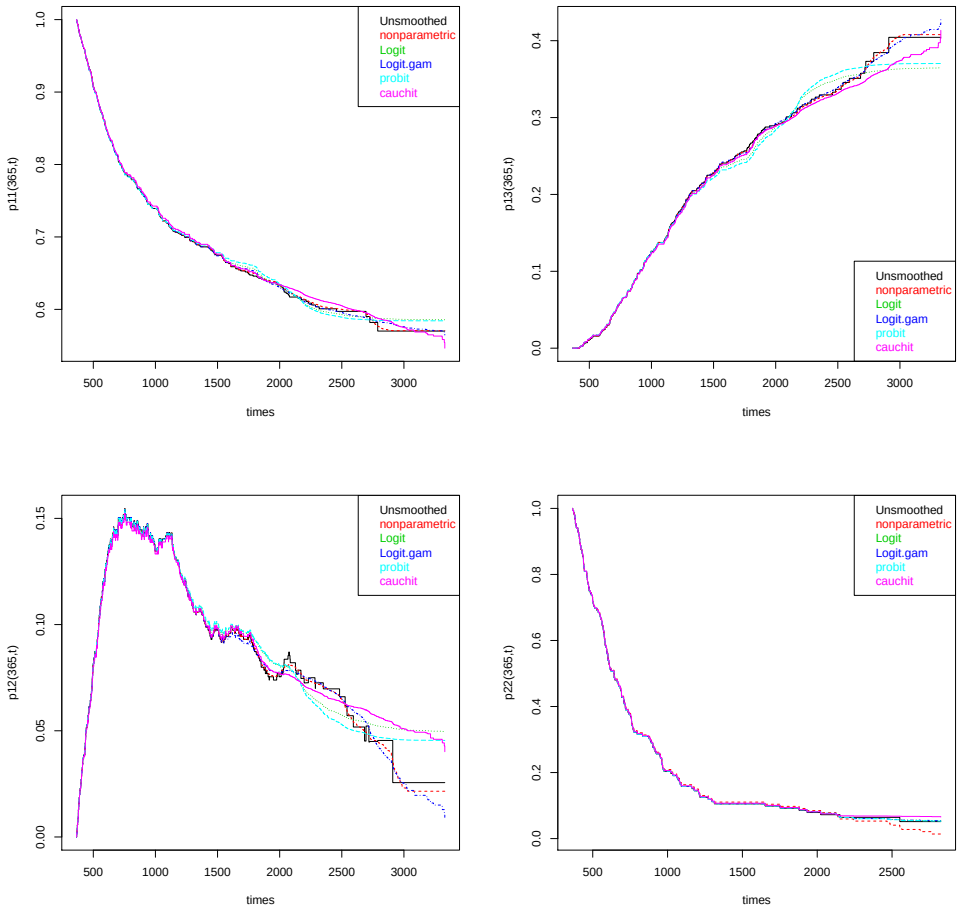


Figure 2: Estimated transition probabilities.

References

[1] Aalen, O.O. and Johansen, S. (1978). An empirical transition matrix for non homogeneous Markov and chains based on censored observations Matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* 5, 141–150.

[2] Cao, R., Lopez-de Ullibarri, I., Janssen, P., and Veraverbeke, N. (2005). Presmoothed Kaplan-Meier and Nelson-Aalen estimators. *Journal of Nonparametric Statistics* 17, 31–56.

[3] de Uña-Álvarez, J. and Meira-Machado, L. (2015). Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study. *Biometrics* 71, 364–375.

[4] Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley & Sons, New York.

[5] Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53, 457–481.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

ON THE PARAMETERS ESTIMATION OF HIV DYNAMIC MODELS

Diana Rocha[1], Sónia Gouveia[2,1], Carla Pinto[3,4], Manuel Scotto[5], João Nuno Tavares[3], Emília Valadas[6], Luís Filipe Caldeira[6]

[1] Center for R&D in Mathematics and Applications - CIDMA, University of Aveiro (diana.isa.rocha@ua.pt)
[2] Institute of Electronics and Informatics Engineering of Aveiro - IEETA
[3] Centre of Mathematics of the University of Porto - CMUP
[4] Institute of Engineering of the Polytechnic Institute of Porto - ISEP
[5] CEMAT and Department of Mathematics, IST, University of Lisbon
[6] Infectious Disease Service, Hospital Santa Maria, Lisbon (HSM/CHLN)

ABSTRACT

This work introduces an estimation method to obtain the optimal parameter estimates of a mathematical model from a set of CD4$^+$T values from a HIV patient. To this end, the following scheme is adopted: the first step consists of selecting the candidate with minimum square error, from a set of randomly generated candidates. In the second step, the selected candidate is refined by an optimization algorithm with constraints and bounds (imposed by physiology), resulting on the optimal estimate. The proposed method is illustrated through a simulation study.

Keywords and key sentences: Human Immunodeficiency Virus (HIV), Mathematical Models, Nonlinear Programming, Parameter Estimation.

1. METHODS AND PROCEDURES

This work presents a nonlinear programming approach to estimate the parameters of a HIV dynamic model that mimics the temporal evolution of the clinical markers of a HIV patient. The mathematical model translates known physiological relationships between Viral Load values and the number of CD4$^+$T cells for one HIV infected patient [3, 4], through the following ordinary differential equations:

$$\begin{aligned} \frac{dT(t)}{dt} &= \lambda - d_1 T(t) - (1-\epsilon) k_1 T(t) V(t), \\ \frac{dT^*(t)}{dt} &= (1-\epsilon) k_1 T(t) V(t) - \delta T^*(t), \\ \frac{dV(t)}{dt} &= \pi_1 T^*(t) - c V(t), \end{aligned} \tag{1}$$

where the state variables are the viral load $V(t)$ and the number of CD4$^+$T cells defined as $CD4(t) = T(t) + T^*(t)$, that is, the number of uninfected and infected CD4$^+$T cells.

Furthermore, we denote as $(T(0), T^*(0), V(0)) = (T_0, T_0^*, V_0)$ - the initial conditions of the model. This model also incorporates parameters with clinical interpretation namely $\boldsymbol{\theta} = (d_1, \epsilon, k_1, \delta, \pi_1, c)$, where d_1 is the difference between loss from cell death and gain due to cell division (and $\lambda = d_1 T_0$ is the proliferation rate of uninfected target cells), $0 \leq \epsilon \leq 1$ denotes the effectiveness of therapy, k_1 is the infectivity rate, δ is the death rate of infected cells, π_1 is the average number of virions produced by a single infected cell and c is the clearance rate of free virions.
Let $CD4(t_i)$ be the observed number of CD4$^+$T cells at time $t_i, i = 1, 2, ..., n$, and define $\widehat{CD4}(t_i) = T(t_i) + T^*(t_i)$, as the estimate of $CD4(t_i)$ provided by the model (1). The *optimal* parameter estimates, say $\widehat{\boldsymbol{\theta}}$, can be obtained by minimizing the square error between model estimates and observed CD4 values. Thus, the nonlinear programming algorithm can be formulated as

$$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\widehat{CD4}(t_i) - CD4(t_i))^2 = \sum_{i=1}^{n} e_{t_i}^2 \\ \text{subject to} \quad & \sum_{i=1}^{n} e_{t_i} = 0 \\ \text{and} \quad & lb \leq \boldsymbol{\theta} \leq ub \end{aligned}$$

where the restriction guarantees that the numerical solution $\widehat{\boldsymbol{\theta}}$ verifies that property of the minimum least square method (i.e. equal contribution of negative and positive deviations from observations). Finally, $\widehat{\boldsymbol{\theta}}$ is restricted to physiological lower and upper bounds, respectively $lb = (0.01, 0, 10^{-11}, 0.24, 50, 2.39)$ and $ub = (0.02, 1, 10^{-5}, 0.7, 10000, 23)$ [2, 4]. The optimization procedure was implemented with MATLAB$^{\text{TM}}$ function $fmincon$, that starts at an initial solution $\boldsymbol{\theta}^*$ to find a minimizer $\widehat{\boldsymbol{\theta}}$ of $f(\boldsymbol{\theta})$ subject to the above-mentioned restrictions and bounds. The initial solution $\boldsymbol{\theta}^*$ is obtained as that minimizing $f(\boldsymbol{\theta})$ in a set of 1000 candidates randomly generated from a multivariate uniform distribution on lb and ub.

The HIV dynamic model (1) was implemented with MATLAB$^{\text{TM}}$ function *ode*45. This function uses an explicit Runge-Kutta formula, namely the Dormand-Prince pair [1], that computes the solution at time t_k based on the solution at time t_{k-1}. Furthermore, when the integration is considered in a time span, the algorithm runs with a variable time step for efficient computation. In this case, temporal resampling is needed to obtain the solutions at specific $t_i, i = 1, 2, ..., n$ (continuous time). Alternatively, the solver can provide the solution at requested time points t_i with its own built-in interpolation algorithm (discrete time). The differences between continuous/discrete time solutions were used to determine if differences between solutions at $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$, as described above, are actually relevant.

2. RESULTS ANALYSIS

The estimation procedure was illustrated through a simulation study. Similarly spaced $CD4(t)$ and $V(t)$ observations were obtained within the interval $[0, 120]$(days), reproducing the temporal evolution of one patient where $\boldsymbol{\theta}_0 = (0.012, 0, 0.75 \times 10^{-6}, 0.39, 790, 3)$ from the beginning of the HIV infection (by setting $(T_0, T_0^*, V_0) = (11 \times 10^3, 0, 10^{-6})$ i.e. a large initial number of uninfected cells T_0 and low values for the initial number of infected cells T_0^* and viral load V_0) and undergoing no HIV treatment ($\epsilon = 0$) [4]. Within this setting, we obtain a set of $n = 18$ observations representing the temporal trajectory of the patient in a clinical follow-up every 7 days ($t_i = [0, 7, 14, 21, 28, \ldots, 119]$, e.g. $t_5 = 28$). Afterwards, 100 replicas of that trajectory were randomly generated, by adding an error to the CD4$^+$T values, in accordance with the fact that laboratory CD4$^+$T measurements have an error of about

20% of the measured value (i.e. $e \sim N(0, \boldsymbol{\sigma}_e^2)$) [5]. Note that the quadratic deviation (of the realizations) of e from zero is such that $\sum_{i=1}^{n} e_{t_i}^2 = f(\boldsymbol{\theta}_0) \approx \boldsymbol{\sigma}_e^2(n-1)$ as $\boldsymbol{\theta}_0$ is the simulation reference. For each replica, we obtain $\widehat{\boldsymbol{\theta}}_0$ as the solution of the optimization problem.

Figure 1(a) shows the distribution of $f(\boldsymbol{\theta}_0)$ for the replicas and, being the distribution centred around the simulation reference, as expected. Figures 1(b–c) display the CD4 trajectory lines obtained from $\boldsymbol{\theta}_0$ and from $\widehat{\boldsymbol{\theta}}_0$ for two different simulated replicas. The curves in Figure 1(b) are quite similar as they correspond to a replica with $f(\boldsymbol{\theta}_0)$ close to the simulation reference. For replicas with $f(\boldsymbol{\theta}_0)$ higher than the simulation reference, as the one in Figure 1(c), the improvement of fit from $\boldsymbol{\theta}_0$ to $\widehat{\boldsymbol{\theta}}_0$ is relevant, as $\widehat{\boldsymbol{\theta}}_0$ produces a curve which is clearly more adjusted to the simulated data than that obtained with $\boldsymbol{\theta}_0$. Figure 1(c) also suggests that the observations do not contribute equally to the performance increase: e.g. residuals at high derivative values (black dots) are increased for $f(\boldsymbol{\theta}_0)$ and reduced when $\boldsymbol{\theta}_0$ is replaced by $\widehat{\boldsymbol{\theta}}_0$.

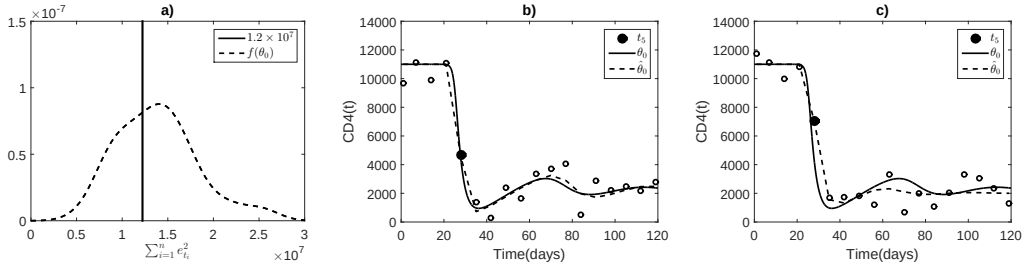


Figure 1: **(a)** Distribution of $f(\boldsymbol{\theta}_0)$ evaluated for 100 replicas (i.e. $\boldsymbol{s}_{\hat{e}}^2(n-1)$ where $\hat{e}$ are the residuals of the model with parameters $\boldsymbol{\theta}_0$, for each replica). The vertical line locates $\boldsymbol{\sigma}_e^2(n-1) = 1.2 \times 10^7$ used in the simulation. **(b–c)** CD4 trajectory line from $\boldsymbol{\theta}_0$ and optimized $\widehat{\boldsymbol{\theta}}_0$ for two different replicas: (b) $\boldsymbol{s}_{\hat{e}}^2(n-1) \approx 1.2 \times 10^7$ and (c) $\boldsymbol{s}_{\hat{e}}^2(n-1) \approx 2 \times 10^7$. The circles represent the simulated observations and the black dot highlights time t_5.

Figure 2 compares the modelling results with $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_0$. As observed in Figures 2(a–b), the distribution of $f(\hat{\boldsymbol{\theta}}_0)$ is more shifted towards the small deviations than $f(\boldsymbol{\theta}_0)$ and $f(\boldsymbol{\theta}_0) - f(\hat{\boldsymbol{\theta}}_0)$ is positive for almost all replicas, thus evidencing that lower squared errors are achieved for $\hat{\boldsymbol{\theta}}_0$. Also, as illustrated in Figures 2(c–d), the $f(\boldsymbol{\theta}_0) - f(\hat{\boldsymbol{\theta}}_0)$ differences are higher than those obtained by choosing continuous/discrete time option for the model numerical resolution. This suggests that differences between $f(\boldsymbol{\theta}_0)$ and $f(\hat{\boldsymbol{\theta}}_0)$ are indeed relevant.

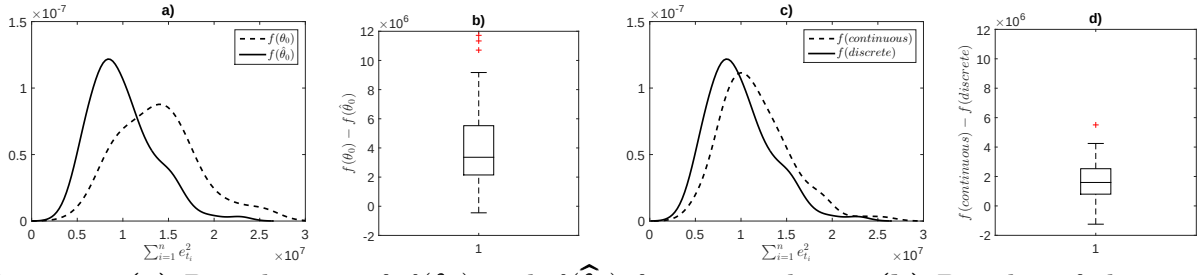


Figure 2: **(a)** Distribution of $f(\boldsymbol{\theta}_0)$ and $f(\widehat{\boldsymbol{\theta}}_0)$ for 100 replicas. **(b)** Boxplot of the paired differences $f(\boldsymbol{\theta}_0) - f(\hat{\boldsymbol{\theta}}_0)$. **(c–d)** Same plots as before for $f(\widehat{\boldsymbol{\theta}}_0)$ and continuous/discrete time.

Figure 3(a) shows the association between performance increase of $\hat{\boldsymbol{\theta}}_0$ with respect to $\boldsymbol{\theta}_0$, as measured by $f(\boldsymbol{\theta}_0) - f(\hat{\boldsymbol{\theta}}_0)$, and the dispersion of the residuals introduced in the simulation process. The correlation turns out to be moderate ($r = 0.60$). The effect of the residual at each time t_i was further investigated, by computing the correlation between $f(\boldsymbol{\theta}_0) - f(\hat{\boldsymbol{\theta}}_0)$ and the squared residual value at time t_i. Figure 3(b) shows a high correlation between

$e_{t_5}^2$ and performance increase ($r = 0.91$), where higher $e_{t_5}^2$ values are associated with higher performance improvement. Furthermore, note that a large part of the residuals dispersion is due to the contribution of e_{t_5}. This analysis suggests that the observations do not contribute equally to the performance increase. In this case, t_5 corresponds to the time point with the largest residual values for $\boldsymbol{\theta}_0$ and highest derivate in the CD4 curve (Figures 1(b–c)).

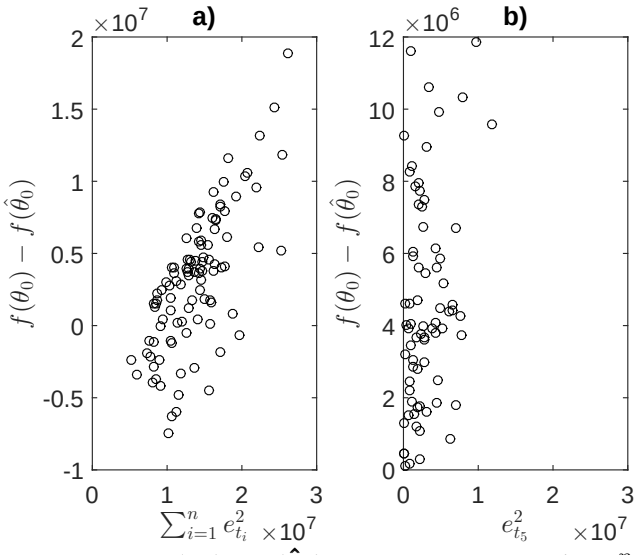


Figure 3: Dispersion diagram of $f(\boldsymbol{\theta}_0) - f(\hat{\boldsymbol{\theta}}_0)$ as a function of **a)** $\sum_{i=1}^{n} e_{t_i}^2$ and **b)** $e_{t_5}^2$, the time that maximizes correlation between $f(\boldsymbol{\theta}_0) - f(\hat{\boldsymbol{\theta}}_0)$ and $t_i, i = 1, 2, ..., n$. For the remaining time points the correlation was $< |0.20|$. Each dot represents one of the 100 simulation replicas.

This work addresses the problem of estimating the parameters of a HIV dynamic model from a set of observations. The proposed method is validated via a data simulated with reference parameters $\boldsymbol{\theta}_0$. The results indicate that the replacement of $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}_0$ decrease the fit error in a value that is greater than the continuous/discrete reduction factor. Therefore, the performance increase is relevant. Moreover, the algorithm provides adequate $\hat{\boldsymbol{\theta}}_0$ estimates, thus foreseeing promising results in real clinical data.

Acknowledgements

This work was partially funded by the Foundation for Science and Technology (FCT), through national funds (MEC) and european structural (FEDER), through the UID/MAT/04106/2013 (CIDMA), UID/CEC/00127/2013 (IEETA) and UID/MAT/00144/2013 (CMUP) projects. Diana Rocha acknowledges the FCT grant (ref. SFRH/BD/107889/2015).

References

[1] Dormand, J.R., Prince, P.J. (1980) A Family of Embedded Runge-Kutta Formulae. *J. Comp. Appl. Math.*, 6, 19–26.

[2] Hadjiandreou, M.M., Conejeros, R., Wilson, D.I. (2009) Long-term HIV Dynamics Subject to Continuous Therapy and Structured Treatment Interruptions. *Chem. Eng. Sci.*, 64, 1600–1617.

[3] Nowak M.A., May R.M. (2000) Virus Dynamics: Mathematical Principles of Immunology and Virology. Oxford: Oxford University Press.

[4] Stafford, M.A., Coreya L., Caob Y., Daardd E.S., Hob D.D., Perelson A.S. (2000) Modeling Plasma Virus Concentration During Primary HIV Infection. *J. Theor. Biol.*, 203, 285–301.

[5] Whitby, L. et al (2013) Comparison of Methodological Data Measurement Limits in CD4$^+$T Lymphocyte Flow Cytometric Enumeration and Their Clinical Impact on HIV Management. *Cytometry Part B (Clinical Cytometry)*, 84B, 248–254.

Comunicações Orais

III Encontro Luso-Galaico de Biometría
Aveiro, 28 - 30 June 2018

AN APPLICATION OF STRATIFIED BOOTSTRAP IN THE DETERMINATION OF LIPID AND LIPOPROTEIN REFERENCE PERCENTILES FOR THE PORTUGUESE POPULATION

Cibelle Mariano [1,2], Marília Antunes [3,4] e Mafalda Bourbon [1,2]

[1]Cardiovascular Research Group, Research and Development Unit, Department of Health Promotion and Chronic Diseases, National Institute of Health Doutor Ricardo Jorge, Lisbon, Portugal;
[2]Biosystems & Integrative Sciences Institute – BioISI, Faculty of Sciences, University of Lisbon, Lisbon, Portugal
[3]Department of Statistics and Operations Research, Faculty of Sciences, University of Lisbon, Lisbon, Portugal;
[4]Centre of Statistics and its Applications – CEAUL, Faculty of Sciences, University of Lisbon, Lisbon, Portugal.

ABSTRACT

The establishment of population specific, age and gender, reference intervals are recommended for a better interpretation of clinical laboratory tests and for patient care. There are different ways to establish reference values, and percentiles estimation is one of them. Lipids and lipoproteins reference values based on population specific percentiles were never determined for the Portuguese population. The aim of this study was to determine lipid and lipoprotein percentiles for the Portuguese population and to compare it with other population studies. A total of 866 individuals were included for analysis. The 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles were obtained for total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), apolipoprotein A1 (apoA1), and apolipoprotein B (apoB) (Table 1). The e_COR sampling design allowed the powerful estimation of the national prevalence of cardiovascular risk factors and hence sample size was calculated with this purpose. As a consequence, the total sample was not representative of the Portuguese population regarding age and gender distribution and could not be used directly to estimate percentiles of the parameters of interest. To overcome this, a stratified Bootstrap approach was chosen. It consisted of resampling, from the total sample, a high number of subsamples following a sampling scheme that respected both age and gender distribution of the Portuguese population across the regions. The estimated percentiles were then compared with the same percentiles from other

populations, by plotting the percentile graphs from each study together with the estimated percentiles and corresponding estimated 95% confidence intervals. Heat colour matrices were further constructed for a more general overview and comparison of percentiles between studies. We provided for the first time reference values for lipid biomarkers for the Portuguese population, based on lipid percentiles. Finally, we also showed a very visual and feasible method for comparison analysis of the percentile values. We strongly encourage the estimation of population-specific reference values, for the definition of normal and at risk values.

Keywords and key sentences: Dyslipidaemia, Lipid biomarkers, Lipid percentiles, Reference values, at risk values, Stratified bootstrap.

1. INTRODUCTION

Reference values of plasma biomarkers are statistically derived numbers from a reference population. The individuals should be randomly selected from a reference population, ideally using specific criteria (age, gender, race, etc.) and including exclusion criteria (e.g. tobacco use, medications, etc.)[1]. Well-established biomarker reference ranges provide a baseline to assess the clinical status of an individual and/or population. These biomarkers are commonly used in basic and clinical research and in community assessments, such as policymakers use population-level biomarkers for screening, surveillance, and monitoring/evaluation of interventions. Clinicians use biomarkers mainly for diagnosis, prognosis, and treatment [2].

There are different ways to establish reference values, and percentiles estimation is one of them. Lipids and lipoproteins percentiles were never determined for the Portuguese population, although they have been for specific subpopulations as part of different studies. In 2013, Cortez-Dias and colleagues determined TC, LDL-C, HDL-C, and TG percentiles for a specific Portuguese population, primary health care users, but this has limited application due to sample bias [3].

Biomarker percentiles are of extreme importance for the definition of normal ranges intervals, being useful for giving the relative standing of an individual in a population. They are essentially the rank position of an individual. The percentiles calculation has the advantage that these are not strongly influenced by extreme values of the distribution (as the mean value), and do not requires normally distributed data, which means that can be calculated, even if the data are skewed [4].

Percentiles can be obtained by different strategies, including bootstrap methods that are increasingly being used in the medical literature, especially for non-Gaussian population's distributions or in the absence of any knowledge of a distribution. In a bootstrap, a set of data is randomly resampled with replacement, multiple times, and statistical conclusions are drawn from the data collection. The nonparametric bootstrap is a very computer-intensive method, but with a valuable application in the determination of confidence intervals of a quantile (e.g. 0.05 to 0.95) or percentile (e.g. 5th to 95th) [5,6]. In the presence of a number of samples of the several strata of a population which, as a whole constitutes a non representative sample of the population, stratified bootstrap is a way of obtaining good estimates for population quantiles.

Here, we provide for the first time the percentiles for lipid metabolism biomarkers of the Portuguese population, namely TC, LDL-C, HDL-C, TG, apolipoprotein B (apoB),

apolipoprotein A1 (apoA1), small dense LDL-C (sdLDL-C), lipoprotein(a) [Lp(a)], and also for non-HDL-C, apoB/apoA1 and sdLDL-C/LDL-C ratios, and remnant cholesterol, for the Portuguese population. We then compared these percentiles with those from a Portuguese primary care study [4] and with a Spanish [7] and American populations studies [8,9].

2. RESULTS

A schematic representation of the procedure adopted to assess the homogeneity distribution of lipid parameters among regions can be found in Figure 1. The homogeneity of the distribution of lipid parameters among regions was tested within each age group and gender using Kruskal-Wallis non-parametric statistical test. For the age groups with evidence of lack of homogeneity, a Kolmogorov-Smirnov test was applied between regions to assess lack of homogeneity among pairs of regions. Regions for which the homogeneity hypothesis was not rejected were grouped and analysed as one individual stratum. Considering for each region/group of homogeneous regions, the respective stratum weights percentiles and 95% confidence intervals (CI) were estimated by stratified bootstrapping. Each combined bootstrap sample was obtained by sampling with replacement from each stratum a number of observations proportional to the stratum weight in the population.

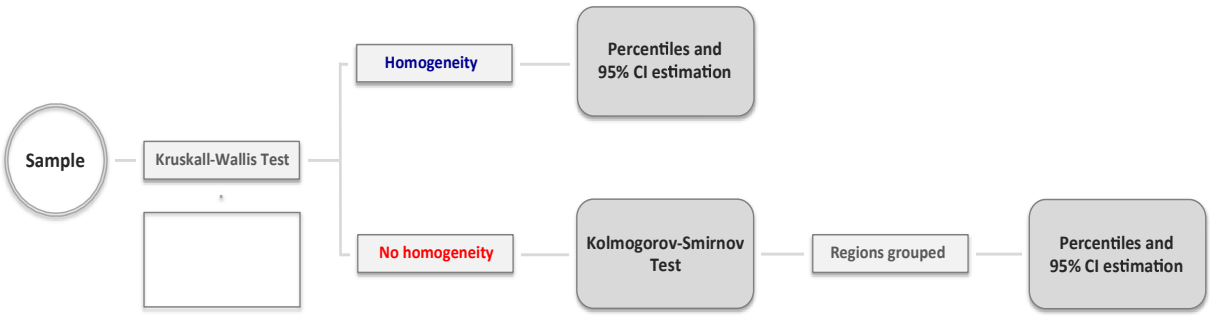


Figure 1: Homogeneity assessment scheme.

3. CONCLUSIONS

Taking all results into consideration, we recommend the newly determined lipid percentiles of the Portuguese population to be used in a clinical context. All percentiles were estimated using stratified bootstrap methodology, taking into account gender and age-specific stratum weights, which were used to overcome the limitation of the e_COR sample not being representative of our population due to the study design. This way, the values obtained are representative of the Portuguese population.

The 50th percentile can be considered normal and/or a range between 25th percentile and 75th percentile could be adopted as a normal reference interval. Above the 90th percentile for TC, LDL C, TG, apoB, sdLDL-C, Lp(a), non-HDL-C apoB/apoA1 and remnant cholesterol, or below the 10th percentile for HDL and apoA1 can be considered at risk. High risk can be defined above the 95th or below the 5th percentiles, and so it can also be defined as the cut off for the different lipid disorders. It is important to note that these values were estimated based on what could be considered a general population with untreated lipid values. Using these

percentile values as reference, will provide a picture of how deviated an individual patient's value is from the expected in the global population. These newly determined reference values for lipid biomarkers will allow a correct dyslipidaemia assessment and the use of these reference values in the clinic, for a better patient care and management.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Dr. Ana Catarina Alves for coordinating the e_COR field work, Mrs. Marta Alvim and Mrs. Ana Raimundo for coordinating the biochemical determinations and Mrs Ana Raimundo also for technical assistance with sdLDL-C and Lp(a) biochemical determinations, and all technicians of the "Unidade de Diagnóstico e Referência" (DPSPDNT – INSA) for performing the lipid profile. CM was supported by a PhD student grant [SFRH/BD/52494/2014]. The work of Marília Antunes was supported by national funds through FCT under the project UID/MAT/00006/2013.

References

[1] Horowitz, G. L. (2008). Reference Intervals: Practical Aspects. *EJIFCC*, 19 (2), 95–105.

[2] Schulte, P. A. (2005). The Use of Biomarkers in Surveillance, Medical Screening, and Intervention. *Mutat. Res. Mol. Mech. Mutagen*, 592 (1–2), 155–163.

[3] Cortez-Dias, N.; Robalo Martins, S.; Belo, A.; Fiúza, M. (2013). Characterization of Lipid Profile in Primary Health Care Users in Portugal. *Rev. Port. Cardiol*., 32 (12), 987–996.

[4] Altman, D. Practical Statistics for Medical Research; London: Chapman and Hall, 1991.

[5] Henderson, A. R. (2005). The Bootstrap: A Technique for Data-Driven Statistics. Using Computer-Intensive Analyses to Explore Experimental Data. *Clin. Chim. Acta*, 359 (1), 1-26.

[6] Desharnais, B.; Camirand-Lemyre, F.; Mireault, P.; Skinner, C. D. (2015). Determination of Confidence Intervals in Non-Normal Data: Application of the Bootstrap to Cocaine Concentration in Femoral Blood. *J. Anal. Toxicol*., 39 (2), 113–117.

[7] Gómez-Gerique, J.A.; Gutiérrez-Fuentes, J.A.; Montoya, M.T.; Porres, A.; Rueda, A.; Avellaneda, A.; Rubio, M. Á. (1999). Perfil Lipídico de La Población Española: Estudio DRECE (Dieta Y Riesgo de Enfermedad Cardiovascular En España). *Med. Clin*., 113 (19), 730–735.

[8] Contois, J. H.; McNamara, J. R.; Lammi-Keefe, C. J.; Wilson, P. W.; Massov, T.; Schaefer, E. J. (1996). Reference Intervals for Plasma Apolipoprotein B Determined with a Standardized Commercial Immunoturbidimetric Assay: Results from the Framingham Offspring Study. *Clin. Chem*., 42 (4), 515–523.

[9] Bachorik, P. S.; Lovejoy, K. L.; Carroll, M. D.; Johnson, C. L. (1997). Apolipoprotein B and AI Distributions in the United States, 1988-1991: Results of the National Health and Nutrition Examination Survey III (NHANES III). *Clin. Chem*., 43 (12), 2364–2378.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

HOW ASYMMETRIC IS VOLATILITY IN HRV?

Argentina Leite[1], Ana Paula Rocha[2], Maria Eduarda Silva[3]

[1]Escola de Ciências e Tecnologia, Universidade de Trás-os-Montes e Alto Douro & C-BER & INESC TEC, Portugal tinucha@utad.pt
[2]Faculdade de Ciências, Universidade do Porto & CMUP, Portugal aprocha@fc.up.pt
[3]Faculdade de Economia, Universidade do Porto & CIDMA, Portugal mesilva@fep.up.pt

ABSTRACT

The analysis of Heart Rate Variability (HRV) has proved important to assess the integrity of the cardiovascular regulatory system and several approaches and methodologies to analyse HRV may be found in the literature. This work considers asymmetric GARCH type models to provide further non linear characterization of HRV.

Keywords and key sentences: HRV, ARFIMA, EGARCH, GJRGARCH.

1. INTRODUCTION

Heart Rate Variability (HRV) data display non stationary characteristics and exhibit long memory and time-varying conditional variance (usually designated by volatility) which may contain indicators of current disease or warnings about impending diseases. Traditionally, HRV data can be characterized by linear AutoRegressive (AR) models, which describe only short memory in the mean. These models combined with recursive least squares have been used to estimate the volatility in HRV data. An alternative approach to describe the dynamics in HRV data was proposed by Leite *et al* [1] using Fractionally Integrated AutoRegressive Moving Average (ARFIMA) models with heteroscedastic errors. These models which are an extension of the AR models usual in the analysis of HRV and may be used to capture and remove long memory and estimate the volatility in 24 hour HRV recordings [1], satisfy the following equations

$$\begin{aligned} \phi(B)(1-B)^d x_t &= \epsilon_t && (1)\\ \epsilon_t &= \sigma_t z_t && (2)\\ \sigma_t^2 &= \mathrm{Var}(\epsilon_t | H_{t-1}) && (3) \end{aligned}$$

where B is the backward-shift operator. Equation (1) describes the conditional mean of the process with serially uncorrelated residuals ϵ_t and is said an ARFIMA$(p, d, 0)$ where d, the long-memory parameter, determines the long-term behaviour in mean. Equations (2) and (3) describe the conditional variance of the process which varies over time as in time-varying AR

models. In (2), ϵ_t are called shocks and z_t, independent and identically distributed random variables with zero mean and unit variance, are the standardized shocks.
There are several models to govern the evolution of σ_t^2, [2]. The most common is the Generalized Autoregressive Conditionally Heteroscedastic, denoted by GARCH$(1,1)$, model under which

$$\sigma_t^2 = u_0 + v_1\,\sigma_{t-1}^2 + u_1\,\epsilon_{t-1}^2 \tag{4}$$

where $u_0 > 0$, v_1, $u_1 \geq 0$, $v_1 + u_1 < 1$. The parameters, u_1 and v_1 characterize the volatility clustering phenomena observed in many data sets. The persistence parameter (amount of volatility clustering captured by the model) for this model is $v_1 + u_1$. A further characteristic present in many data sets is the so called asymmetric effect, which is not captured by the GARCH model but may be modeled by the EGARCH$(1,1)$ proposed by [3] and the GJRGARCH$(1,1)$ proposed by [4]. The former models the dynamics of the conditional volatility 3 by

$$\log\sigma_t^2 = u^* + v_1\log\sigma_{t-1}^2 + u_1\,|z_{t-1}| + \xi_1 z_{t-1} \tag{5}$$

with $u^* = u_0 - u_1\sqrt{\frac{2}{\pi}}$ and $z_t = \epsilon_t/\sigma_t$. The parameters u_1 and v_1 characterise the volatility clustering phenomena and the parameter ξ_1 describes the leverage effect. The persistence parameter for this model is v_1. The impact of positive shocks, $\epsilon_{t-1} > 0$, is $(u_1 + \xi_1)\frac{\epsilon_{t-1}}{\sigma_{t-1}}$, while for negative shocks it is $(u_1 - \xi_1)\frac{\epsilon_{t-1}}{\sigma_{t-1}}$. If $\xi_1 = 0$, $\log\sigma_t^2$ responds symmetrically to ϵ_{t-1}. The GJRGARCH$(1,1)$ models the dynamics of volatity 3 as:

$$\sigma_t^2 = u_0 + v_1\,\sigma_{t-1}^2 + u_1\,\epsilon_{t-1}^2 + \xi_1\,I_{t-1}\,\epsilon_{t-1}^2 \tag{6}$$

where $u_0 > 0$, v_1, $u_1 \geq 0$, $u_1 + \xi_1 \geq 0$, $v_1 + u_1 + \xi_1/2 < 1$ and the indicator function $I_{t-1} = 1$ if $\epsilon_{t-1} < 0$ and 0 otherwise. The parameter ξ_1 describes the leverage effect and the persistence parameter for this model is $v_1 + u_1 + \xi_1/2$. A positive shock ϵ_{t-1} contributes $u_1\epsilon_{t-1}^2$ to σ_t^2, whereas a negative shock ϵ_{t-1} has a larger impact $(u_1 + \xi_1)\epsilon_{t-1}^2$ with $\xi > 0$. The GARCH model is nested in the GJRGARCH model. If the leverage coefficient is zero, then the GJRGARCH model reduces to the GARCH model.
In this work, we apply the two prominent asymmetric volatility models, ARFIMA$(p,d,0)$-EGARCH$(1,1)$ and ARFIMA$(p,d,0)$-GJRGARCH$(1,1)$ models, to 24 hour HRV recordings provided by PhysioNet [5]: five from normal subjects, N, five from heart failure, C, and five from atrial fibrillation patients, A (http://www.physionet.org/challenge/chaos/).

2. ASYMMETRIC MODELING OF HRV

The description of 24 hours HRV data (long recordings, approximately 100000 beats) is achieved by asymmetric modeling combined with adaptive segmentation [1]: long records are decomposed into short records of variable length and the break points are identified by Akaike Information Criterion. The short records thus obtained have a minimum length 512 and are subsequently modeled using ARFIMA-EGARCH and ARFIMA-GJRGARCH models.
The HRV data are analysed in three periods: 6 hours during day, 6 hours during night and the whole 24 hours. First, McLeod-Li test is applied to the residuals of ARFIMA model $(\hat{\epsilon}_t)$ and the percentage of segments with conditional heteroscedasticity are reported in Table 1. The results indicate that not only the conditional heteroscedasticity is characteristic of groups N and C but also the percentage of segments with conditional heteroscedasticity is higher during the night period.
Next, the presence of asymmetric effects in the conditional volatility is investigated by fitting EGARCH and GJRGARCH models to the HRV data. A simple diagnostic for uncovering asymmetric effects is to test the significance to the parameter ξ_1.The percentage of segments

with asymmetric effect is summarized in Table 1. The results, similar for both models, indicate that the asymmetric effect is observed in the groups N and C, and preponderant during the night period.

% of Segments with	Period	N	C	A
Heteroscedasticity	24 h	80.7 ± 3.4	80.4 ± 6.4	11.2 ± 4.0
	Day - 6 h	76.2 ± 11.0	81.0 ± 4.9	9.4 ± 10.4
	Night - 6 h	92.9 ± 4.5	86.8 ± 5.7	6.1 ± 5.8
Asymmetric effect	24 h	60.8 ± 10.7	54.1 ± 17.7	5.5 ± 2.4
	Day - 6 h	58.0 ± 14.4	51.2 ± 19.7	1.7 ± 2.6
EGARCH model	Night - 6 h	67.9 ± 18.8	69.4 ± 17.5	5.4 ± 6.4
Asymmetric effect	24 h	57.0 ± 15.0	43.6 ± 14.4	4.4 ± 2.7
	Day - 6 h	53.4 ± 17.2	42.2 ± 11.7	3.5 ± 6.5
GJRGARCH model	Night - 6 h	67.4 ± 21.6	49.2 ± 20.6	3.5 ± 6.7

Table 1: Percentage of segments (mean $\pm$ sd) with conditional heteroscedasticity and with asymmetric effect for the three groups of patients: normal subjects N, congestive heart failure patients C and patients undergoing atrial fibrillation A, during 24 hours, 6 hours of night and 6 hours of day periods.

Furthermore, it is interesting to assess and compare the HRV asymmetric response in volatility under the two models, (5) and (6). To that end, compare the impact of 2 standard deviation negative and positive shocks using the ratio $AFR = \frac{\sigma_t^2(z_{t-1}=-2)}{\sigma_t^2(z_{t-1}=2)}$, hereafter called asymmetric feature ratio. For example a value $AFR = 1.38$ means that the impact of a negative shock of size two standard deviations is about 38% higher than that of a positive shock of the same size,[2].
The evolution of AFR over the 24h for both models illustrated for the healthy subject N1 in Fig. 1(b) and (c), shows circadian variation, with lowest values during the day period.

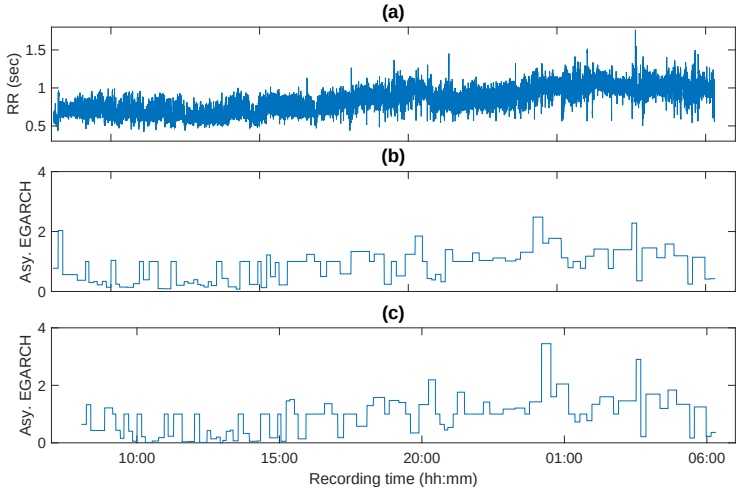


Figure 1: (a) Tachogram of healthy subject-N1, 24 h recordings provided by PhysioNet. Evolution over 24 h of asymmetric features by EGARCH in (b) and GJRGARCH in (c).

The overall results of AF for the 3 groups of patients in the database are presented in Table 2. Within each group, the two models provide similar asymmetric feature. Statistical differences among the three groups of patients, applying the Kruskal-Wallis rank sum test and multiple

Asymmetric feature	Period	**N**	**C**	**A**	$p-$value	N-C	N-A
EGARCH	24h	0.99 ± 0.16	1.45 ± 0.58	1.02 ± 0.02	0.18	—	—
	Day-6h	0.82 ± 0.19	1.25 ± 0.32	1.02 ± 0.02	0.01	$\surd$	—
	Night-6h	1.31 ± 0.08	1.50 ± 0.78	1.04 ± 0.04	0.04	—	$\surd$
GJRGARCH	24h	1.02 ± 0.18	1.19 ± 0.24	1.04 ± 0.03	0.22		—
	Day-6h	0.82 ± 0.25	1.15 ± 0.20	1.01 ± 0.02	0.13	—	—
	Night-6h	1.38 ± 0.28	1.30 ± 0.48	1.08 ± 0.17	0.32	—	—

Table 2: The asymmetric feature (mean $\pm$ st dev) of EGARCH and GJRGARCH models for the three groups of patients provided by PhysioNet: normal subjects N, congestive heart failure patients C and patients were undergoing atrial fibrillation A, during 24 hours, 6 hours of night and 6 hours of day periods.

comparison procedures (5 % level of significance) are reported in Table 2. The results indicate that the asymmetric parameter of EGARCH model differs between the groups N and C during the day period and between the groups N and A during the night period.
In summary, the asymmetric features HRV captured by GARCH type models are promising in differentiating health and disease situations.

ACKNOWLEDGMENT

This work is financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation, COMPETE 2020 Programme (project POCI- 01-0145-FEDER-006961) and ERDF-NORTE2020 (project STRIDE-NORTE-01-0145-FEDER-000033), and by National Funds through the Portuguese funding agency, FCT - Fundacao para a Ciencia e a Tecnologia as part of projects UID/EEA/50014/2013, CIDMA UID/MAT/04106/2013 and CMUP UID/MAT/00144/2013, funded by FCT (Portugal) with national (MEC) and European structural funds through the programs FEDER, under the partnership agreement PT2020.

References

[1] Leite A., Rocha A. P., Silva M. E. (2013). Beyond long memory in heart rate variability: an approach based on fractionally integrated autoregressive moving average time series models with conditional heteroscedasticity. *Chaos*, 23, 023103.

[2] Tsay R. S. (2005). *Analysis of Financial Time Series (2nd ed.)*. Wiley-Interscience.

[3] Nelson D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59, 347-70.

[4] Glosten L. R., Jagannathan R., Runkle D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48, 1779-801.

[5] Goldberger A. L., Amaral L. A. N., Glass L., Hausdorff, J. M., Ivanov P. C., Mark R. G., Mietus J. E., Moody G. B., Peng C. K., Stanley H. E. (2000). PhysioBank, PhysioToolkit and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101, e215-e220.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

NONPARAMETRIC MIXTURE CURE MODELS WITH CURE PARTIALLY KNOWN

M.A. Jácome[1] and I. López-de-Ullibarri[2]

[1]MODES, INIBIC, CITIC, Departamento de Matemáticas, Facultade de Ciencias, Universidade da Coruña, 15071 Coruña, Spain
[2]MODES, INIBIC, CITIC, Departamento de Matemáticas, Escuela Universitaria Politécnica, Universidade da Coruña, 15471, Ferrol, Spain

ABSTRACT

A completely nonparametric mixture cure model is proposed, that allows for some individuals to be identified as cured. The proposed estimators are compared with existing approaches that omit the cure status. The performance of the method is demonstrated in the analysis of a real data set of patients with sarcoma.

Keywords and key sentences: Bandwidth, censored data, cure models, cure indicator, kernel estimation.

1. INTRODUCTION

Typical analysis of time-to-event data assume that all individuals will eventually experience the event of interest. However, when there is evidence of long-term survivors, cure models should be used instead. Mixture cure models, introduced by Boag (1949), are traditionally used. They assume that the population of individuals is made up of two distinct groups: those who will and those who will not experience the event of interest. Hence, the survival function for the population of all individuals, possibly depending on a set of covariates $\mathbf{X}$, is defined as:

$$S(t|\mathbf{x}) = (1 - p(\mathbf{x})) + p(\mathbf{x})S_0(t|\mathbf{x}) \tag{1}$$

where $1-p(\mathbf{x})$ is the probability of cure and $S_0(t|\mathbf{x})$ is the survival function of those individuals experiencing the event or latency. The goal is to estimate the proportion of cured individuals $1 - p(\mathbf{x})$ and the survival function $S_0(t|\mathbf{x})$ for the uncured individuals.
To model $1 - p(\mathbf{x})$, it is common practice to use logistic regression, and the latency $S_0(t|\mathbf{x})$ is usually estimated parametrically (Farewell, 1982; Cantor and Shuster, 1992; Denham et al., 1996, among others) or semiparametrically, using for example the Cox proportional hazards model (Kuk and Chen, 1992; Peng and Dear, 2000; Sy and Taylor, 2000) or the accelerated failure time model (Li and Taylor, 2002; Zhang and Peng, 2007).
Parametric or semiparametric cure models can achieve the greatest efficiency in estimation if their distributional assumptions are satisfied. However, verifying the assumptions is a challenge in practice, and a loss in efficiency and biased estimates might result when the distribution is not correctly identified.

A completely nonparametric approach for both parts of the mixture cure model was first addressed by Maller and Zhou (1992), who proposed a consistent nonparametric estimator of the cure rate, although it could not handle covariates. More recently, Xu and Peng (2014) proposed a nonparametric estimator of $1-p(x)$ which allows for a continuous covariate. López-Cheda et al (2017a, 2017b) extended the existing work by proposing a two-component mixture model with nonparametric forms for both the cure probability and the survival function of the uncured individuals. Although they considered only one covariate, the method can be directly extended to a case with multiple covariates.

The model (1) gives probability of survival of 1 at any time t for cured individuals because $S(t|\mathbf{x}) \to 1 - p(\mathbf{x})$ as $t \to \infty$. For this reason, for a cure individual it is assumed that the lifetime is $T = \infty$ (the event of interest will never take place). A common assumption of traditional cure models is that, due to censoring, cured and uncured subjects cannot be distinguished. Hence, no information about the status of cure is provided, and the indicator of cure is usually modeled as a latent variable. However, this assumption is not entirely valid in some cases, when some information about the cure status is given. One frequent example is the case when individuals are assumed to be cured when the survival time is larger than a given cutoff, known as cure threshold. For instance, 5 years is often used as a threshold when considering cancer recurrence. Other possibility is to infer the cure status from the result of a diagnostic test with a given sensitivity and specificity. A cure threshold has been considered previously by Laska and Meisner (1992) and Tan (2006), who discussed nonparametric estimation without covariates. In the conditional setting, Barajas and Yin (2008) developed a Bayesian approach for estimating a previously unknown and potentially covariate-dependent cure threshold. More recently, Bernhardt (2016) proposed a semiparametric cure rate model with covariates that accommodates different censoring distributions for the cured and uncured groups and also allows for some individuals to be observed as cured when their survival time exceeds a known threshold. For the possibility of knowing the cure status of an individual based on a diagnostic test with a given sensitivity and specificity, see Wu (2010). None of the aforementioned papers deal with a completely nonparametric approach when the cure status is partially known.

Here we propose completely nonparametric method for the estimation of mixture cure models when cure is partially known. The procedure is applied to a cancer dataset of patients with sarcoma.

2. MODEL AND NOTATION

Let Y be the time to the event of interest, C the random censoring time and $\mathbf{X}$ a vector of covariates. The observed lifetime is $T = \min(Y, C)$, and $\delta = \mathbf{1}(Y \leq C)$ is the uncensoring indicator. Suppose that we additionally observe for some individuals the cure indicator ν, where $\nu = 1$ ($\nu = 0$) for cured (uncured) individuals. Note that $\nu = 0$ when $\delta = 1$, since a subject is known to be uncured if the event occurs, and ν is only partially known for the censored times. Thus, the observations are $\{\mathbf{X}_i, T_i, \delta_i, \nu_i\}$, $i = 1, \ldots, n$, and the individuals can be classified into three groups: (a) $\delta_i = 1$ and $\nu_i = 0$, so that the individual is observed to have experienced the event and therefore known to be uncured; (b) $\delta_i = 0$ and $\nu_i = 1$, i.e., the lifetime is censored and the individual is known to be cured; and (c) $\delta_i = 0$ and ν_i is missing, the lifetime is only known to exceed the censoring time and the cure status is unknown. Consider the ordered observed lifetimes $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(n)}$, and $\mathbf{X}_{[i]}$, $\delta_{[i]}$ and $\nu_{[i]}$ the corresponding covariate vector, event indicator and cure indicator concomitants. Without loss of generality, it is assumed that X is a univariate continuous covariate with density function $m(x)$.

Theorem Given $\{X_i, T_i, \delta_i, \nu_i\}$, $i = 1, \ldots, n$, the nonparametric maximum likelihood (NPML) estimators of $1 - p(x)$ and $S_0(t|x)$, in the sense of maximizing the nonparametric mixture

cure model likelihood, are

$$\begin{aligned} & 1-\widehat{p}_h(x) \\ = & \prod_{i=1}^{n}\left(1-\frac{B_{h[i]}(x)}{B_{h[i]}(x)+\sum_{j=i+1}^{n}B_{h[j]}(x)\,\mathbf{1}\left(\nu_{[j]}=0 \text{ or missing}\right)\ +B_C(x)}\right)^{\delta_{[i]}} \end{aligned} \tag{2}$$

where

$$B_{h[i]}(x)=\frac{K_h\left(x-X_{[i]}\right)}{\sum_{j=1}^{n}K_h\left(x-X_j\right)}$$

are the Nadaraya-Watson weights with $K_h(x)=h^{-1}K(x/h)$ the rescaled kernel with bandwidth h, $B_C(x)=\sum_{j=1}^{n}B_{hj}(x)\,\mathbf{1}(\nu_j=1)$ is the sum of the weights of all the cured subjects, and

$$\widehat{S}_{0,b}(t|x)=\frac{\widehat{S}_b(t|x)-(1-\hat{p}_b(x))}{\hat{p}_b(x)}, \tag{3}$$

where $\widehat{S}_b$ is the generalized Kaplan-Meier estimator of the conditional survival function with covariates, proposed by Beran (1981), computed with bandwidth b. The NPML estimator of the probability of cure in (2) has the following properties:

1. **No cures**. If there are no known cures, it reduces to $1-\widehat{p}_h(x)=\widehat{S}_h\left(T^1_{\max}|x\right)$, the nonparametric estimator of the probability of cure proposed by López-Cheda et al (2017a), where $T^1_{\max}$ is the largest uncensored failure time.

2. **No censoring**. If there is no censoring, then $1-\widehat{p}_h(x)=B_C(x)$, the sum of the weights of the known cures.

3. **No covariates**. In an unconditional setting, when an individual is known to be cured only if the lifetime is greater than a known fixed time, then the estimator reduces to the generalized maximum likelihood estimator of the probability of cure in Laska and Meisner (1992).

4. **Common known cure threshold**. If there exists a common known cure threshold $d_i=d$ for $i=1,\ldots,n$, then in the ordered sample the n_1 first observations correspond to individuals with $T_i<d$ either not cured ($\delta_i=1$) or with unknown cure status ($\delta_i=0,\nu_i$ unknown), and the remaining m observations are cured individuals with $T_i\geq d$, $\delta_i=0$ and $\nu_i=1$. In this case, the NPML estimator of $1-p(x)$ reduces to $\widehat{S}_h\left(T^1_{\max}|x\right)$, the nonparametric estimator of the probability of cure proposed by López-Cheda et al (2017a).

The methods are illustrated with a dataset of patients with sarcoma, provided by Angel Díaz-Lagares, a postdoc researcher in Cancer Epigenomics from Translational Medical Oncology (OMT) group, Health Research Institute of Santiago (IDIS) and the University Hospital of Santiago de Compostela (CHUS). The database consists of 261 patients (119 males) aged 20-90 years old. The event is death from sarcoma. A patient is considered cured if he/she will not die from sarcoma no matter how long the study is. A total of 195 observations (74.71%) are censored. Besides clinical covariates such as histological type or tumor depth, each patient is known to be either with tumor or tumor free at the last follow-up. If a patient is tumor free and the survival time exceeds a fixed period of time, set at 5 years in this example, the patient can be diagnosed as cured. The nonparametric estimate of the probability of cure when the cure status is taken into consideration will be given and compared to the estimate obtained if this information is ignored.

ACKNOWLEDGMENT

This research has been supported by MINECO grants MTM2014-52876-R and MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

References

[1] Bernhardt P.W. (2016). A flexible cure rate model with dependent censoring and a known cure threshold. *Statistics in Medicine* 35, 4607–4623.

[2] Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B* 11, 15-53.

[3] Cantor A.B., Shuster J.J. (1992). Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine* 11, 931–937.

[4] Denham J.W., Denham E., Dear K.B., Hudson G.V. (1996). The follicular non-Hodgkin's lymphomas - I. The possibility of cure. *European Journal of Cancer* 32, 470–479.

[5] Farewell, V.T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 38, 1041–1046.

[6] Kuk A.Y.C., Chen C.H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* 79, 531-541.

[7] Laska E.M., Meisner M.J. (1992) Nonparametrics estimation and testing in a cure model. *Biometrics* 48, 1223-1234.

[8] Li C., Taylor J.M.G. (2002). A semi-parametric accelerated failure time cure model. *Statistics in Medicine* 21, 3235–3247.

[9] López-Cheda A., Cao R., Jácome M.A., Van Keilegom, I. (2017a). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Computational Statistics and Data Analysis* 105, 144–165.

[10] López-Cheda A., Jácome M.A., Cao R. (2017b). Nonparametric latency estimation for mixture cure models. *TEST* 26, 353–376.

[11] Maller R.A., Zhou S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika* 79, 731–739.

[12] Nieto-Barajas L.E., Yin G. (2008). Bayesian semiparametric cure rate model with an unknown threshold. *Scandinavian Journal of Statistics* 35, 540-556.

[13] Peng Y., Dear K.B.G. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* 56, 237-243.

[14] Sy J.P., Taylor J.M.G. (2000). Estimation in a Cox proportional hazards models. *Biometrics* 56, 227-236.

[15] Tan F. (2006). *Nonparametric maximum likelihood estimation in cure-rate models based on uncensored and censored data*. Master's Thesis, Concordia University, Portland, Oregon.

[16] Xu J., Peng Y. (2014). Nonparametric cure rate estimation with covariates. *Canadian Journal of Statistics* 42, 1–17.

[17] Wu Y. (2010). *Extension of cure rate model when cured is partially known.* PhD Thesis, University of Medicine and Dentistry of New Jersey.

[18] Zhang M., Peng Y. (2007). A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in Medicine* 26, 3157-3171.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

UM MODELO LINEAR MISTO PARA REGRESSÃO SEGMENTADA LINEAR/QUADRÁTICA

Julio M. Singer[1], Francisco M.M. Rocha[2], Antonio Carlos Pedroso-de-Lima[1], Giovani, L. Silva[3], Giuliana C. Coatti[4] e Mayana Zatz[4]

[1]Departamento de Estatística, Universidade de São Paulo
[2]Escola Paulista de Política, Economia e Negócios, Universidade Federal de São Paulo
[3]Departamento de Matemática, Universidade de Lisboa
[4]Instituto de Biociências, Universidade de São Paulo

RESUMO

Consideramos um estudo realizado com o objetivo de avaliar o efeito da utilização de células tronco para tratamento da esclerose lateral amiotrófica, uma doença neurodegenerativa bastante severa e atualmente sem cura. De maneira semelhante à forma clínica que acomete os pacientes, em camundongos, os sintomas iniciais são tremores nos membros evoluindo progressivamente até a paralisia total com a consequente morte. Um conjunto de animais foi subdividido em três grupos e cada um deles foi tratado semanalmente ou com células mesenquimais ou com pericitos ou com solução de Hank (o veículo utilizado nos demais tratamentos) a partir da oitava semana de vida até a morte natural. Selecionamos uma das quatro respostas para análise e utilizamos modelos mistos para ajuste de regressões segmentadas com a finalidade de avaliar o efeito de tratamento e sexo nas taxas esperadas de variação da resposta ao longo da vida dos animais. Adaptamos um algoritmo desenvolvido por Muggeo et al. [Statistical Modelling (2014)] para ajuste dos modelos e identificamos tópicos para futuras pesquisas.

Palavras chave: efeitos aleatórios, esclerose lateral amiotrófica, pontos de mudança.

1. INTRODUÇÃO

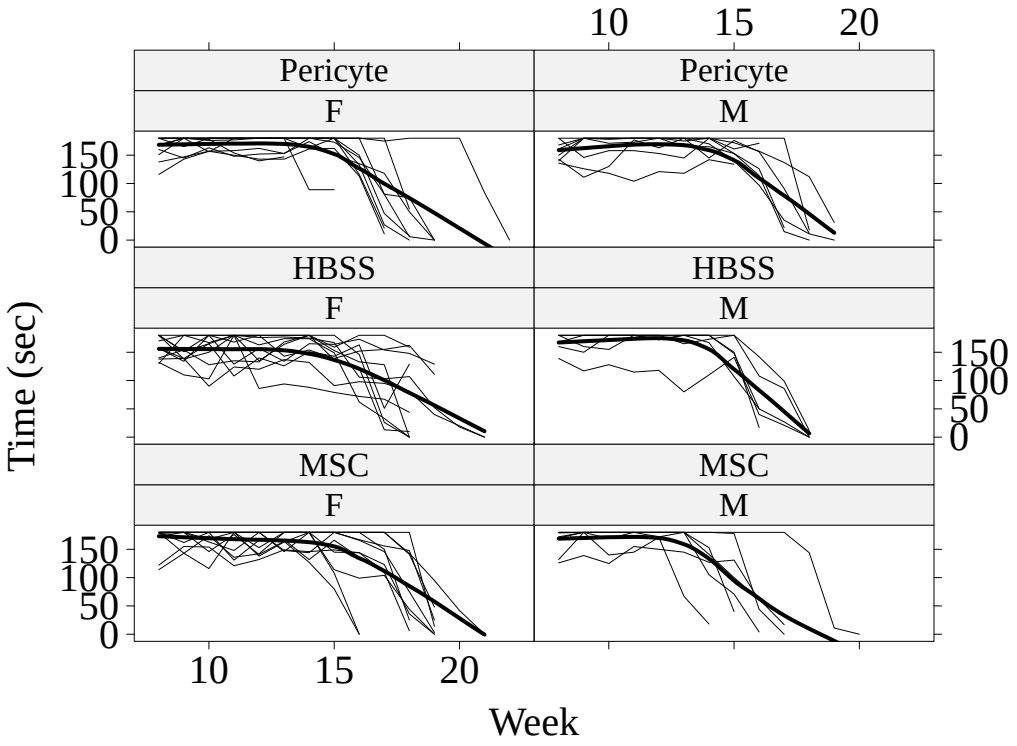
A esclerose lateral amiotrófica (ELA) é uma doença neurodegenerativa fatal causada pela morte de neurônios motores e existência de quadro inflamatório no sistema nervoso central. Dentre os genes identificados para esta doença está o gene SOD1, que codifica uma importante enzima antioxidante humana, a superóxido dismutase 1. Por essa razão, o camundongo transgênico portador de mutação G93A no gene SOD1 é uma importante ferramenta para estudos envolvendo ELA. Uma das abordagens consideradas para a busca de tratamento para esta doença é a terapia com células tronco. As células estromais mesenquimais (mesenchymal stromal cells - MSC), em especial as derivadas de tecido adiposo (adipose-derived stromal cells - ASC), são células que possuem capacidade de se diferenciar em osteócitos, adipócitos,

condrócitos *in vitro*. Embora o seu potencial de diferenciação em células neuronais não tenha sido comprovado, é provável que atuação destas células no tratamento de diversas doenças ocorra por meio da modulação da resposta inflamatória e estresse oxidativo. Apesar dessas características, as MSC constituem uma população bastante heterogênea. Pericitos, por outro lado, representam uma população mais homogênea dado que podem ser obtidos pela técnica de *cell sorting* com marcadores específicos a partir de uma população mista de MSC. Pericitos atuam na manutenção da barreira hematoencefálica, que pode agir na diminuição da aceleração dos sintomas de doenças neurodegenerativas. Nesse contexto, um estudo foi desenvolvido no Instituto de Biociências da Universidade de São Paulo com o objetivo de avaliar o potencial terapêutico de células mesenquimais e de pericitos no camundongo SOD1-G93A. Detalhes podem ser obtidos em Coatti (2015).

2. DESCRIÇÃO DO ESTUDO

Um conjunto de camundongos SOD1-G93A com 8 semanas de vida foi dividido em 3 subconjuntos (11 fêmeas e 8 machos por grupo). Os animais do primeiro subconjunto foram submetidos a injeções semanais com células MSC, aqueles do segundo subconjunto com pericitos e os demais foram submetidos a injeções com HBSS (*Hank's balanced salt solution*), que é o veículo utilizado para as injeções de MSC e pericitos. Os animais foram acompanhados semanalmente até a morte ou paralisia total para análise clínica da progressão da doença por meio de quatro variáveis, nomeadamente, peso, motor score, PaGE (*paw grip endurance*) e *rotarod*. Para exemplificar a análise, selecionamos a variável *rotarod*, que corresponde ao tempo durante o qual o animal permanece no cilindro rotativo de um aparelho *rotarod* (IITC Life Science model 755) com velocidade inicial de 1 rpm constantemente aumentada até a velocidade final de 30 rpm após 180 s. Gráficos de perfis para a variável rotarod com LOESS estão apresentados na Figura 1.

Figura 1: Gráficos de perfis e alisamento LOESS para a variável *rotarod*



3. ANÁLISE ESTATÍSTICA

Uma análise descritiva do comportamento longitudinal da variável *rotarod* sugere que o nível da resposta se mantém estável até um certo instante e que começa a decrescer a partir daí, indicando o início do aparecimento desse sintoma da doença. Além disso, pode-se conjecturar que o comportamento não é uniforme para os seis grupos correspondentes às combinações de tratamento e sexo. Tendo em vista que essas conclusões são compatíveis com a explicação biológica do desenvolvimento da doença, propusemos uma estratégia de análise baseada no ajuste do seguinte modelo misto

$$y_{ijk} = \alpha_{ij} + \{\gamma_{ij}[t_k - \psi_{ij}(\lambda_{ij})]^2\}I(t_k > \psi_{ij}) + e_{ijk} \tag{1}$$

em que y_{ijk} é a resposta do j-ésimo animal observado sob o i-ésimo grupo no k-ésimo instante de avaliação, α_{ij} é o correspondente coeficiente linear da curva que representa o comportamento da resposta antes do ponto de mudança, γ_{ij} é coeficiente do termo quadrático associado à curva que governa a resposta pós ponto de mudança, ψ_{ij}, com $\psi_{ij}(\lambda_{ij}) = [a_1 + a_2\exp(\lambda_{ij})]/[1 + \exp(\lambda_{ij})]$ para restringir o valor de ψ_{ij} ao intervalo (a_1, a_2) em que as respostas são observadas, $\alpha_{ij} = \alpha_i + a_{ij}$, $\gamma_{ij} = \gamma_i + c_{ij}$, $\lambda_{ij} = \lambda_i + \ell_{ij}$ sob a suposição $\mathbf{b}_{ij} = (a_{ij}, c_{ij}, \ell_{ij})^\top \sim N(\mathbf{0}, \mathbf{G})$, $\mathbf{G}$ denota uma matriz de covariâncias (não estruturada) e $e_{ijk} \sim N(0, \sigma^2)$ independente de $\mathbf{b}_{ij}$.
O ajuste é adaptado de Muggeo (2014) e Fasola et al. (2018) e é baseado na expansão de Taylor de

$$f[t_k, \psi_{ij}(\lambda_{ij})] = \gamma_{ij}[t_k - \psi_{ij}(\lambda_{ij})]^2\}I[t_k > \psi_{ij}(\lambda_{ij})]$$

Explicitamente,

$$f[t_k, \psi_{ij}(\lambda_{ij})] \approx f[t_k, \psi_{ij}(\widehat{\lambda}_{ij})] + (\lambda_{ij} - \widehat{\lambda}_{ij})\frac{\partial f[t_k, \psi_{ij}]}{\partial \psi_{ij}}\frac{\partial \psi_{ij}(\lambda_{ij})}{\lambda_{ij}}\Big|_{\lambda_{ij}=\widehat{\lambda}_{ij}}$$

com

$$\frac{\partial f[t_k, \psi_{ij}]}{\partial \psi_{ij}} = h_{ij}(\lambda_{ij}) = -2\gamma_{ij}[t_k - \psi_{ij}(\lambda_{ij})]I[t_k > \psi_{ij}(\lambda_{ij})]$$

e

$$\frac{\partial \psi_{ij}(\lambda_{ij})}{\lambda_{ij}} = g_{ij}(\lambda_{ij}) = \frac{(a_2 - a_1)\exp(\lambda_{ij})}{[1 + \exp(\lambda_{ij}]^2}$$

Consequentemente, pode-se aproximar o modelo (1) como

$$y_{ijk} \approx \alpha_{ij} + f[t_k, \psi_{ij}(\widehat{\lambda}_{ij})] - \widehat{\lambda}_{ij}h_{ij}(\widehat{\lambda}_{ij})g_{ij}(\widehat{\lambda}_{ij}) + \lambda_{ij}h_{ij}(\widehat{\lambda}_{ij})g_{ij}(\widehat{\lambda}_{ij}) + e_{ijk}.$$

Considerando pseudo observações definidas por $y^*_{ijk} = y_{ijk} + \widehat{\lambda}_{ij}h_{ij}(\widehat{\lambda}_{ij})g_{ij}(\widehat{\lambda}_{ij})$, o modelo $y^*_{ijk} = \alpha_{ij} + \lambda_{ij}\widehat{\lambda}_{ij}h_{ij}(\widehat{\lambda}_{ij})g_{ij}(\widehat{\lambda}_{ij}) + e_{ijk}$ sugere o seguinte algoritmo para o ajuste de (1)

1) Fixar $\psi_{ij} = \psi^{(0)}$ e $y^{(0)}_{ijk} = y_{ijk}$.

2) Ajustar o modelo $y^{(0)}_{ijk} = \alpha_{ij} + \gamma_{ij}(t_k - \psi^{(0)})^2 I(t_k > \psi^{(0)}) + e_{ijk}$ para obter $\alpha^{(0)}_{ij}$, $\gamma^{(0)}_{ij}$ e $\lambda^{(0)}_{ij} = \log[(\psi^{(0)} - a_1)/(a_2 - \psi^{(0)})]$.

3) Fixar $r = 1$.

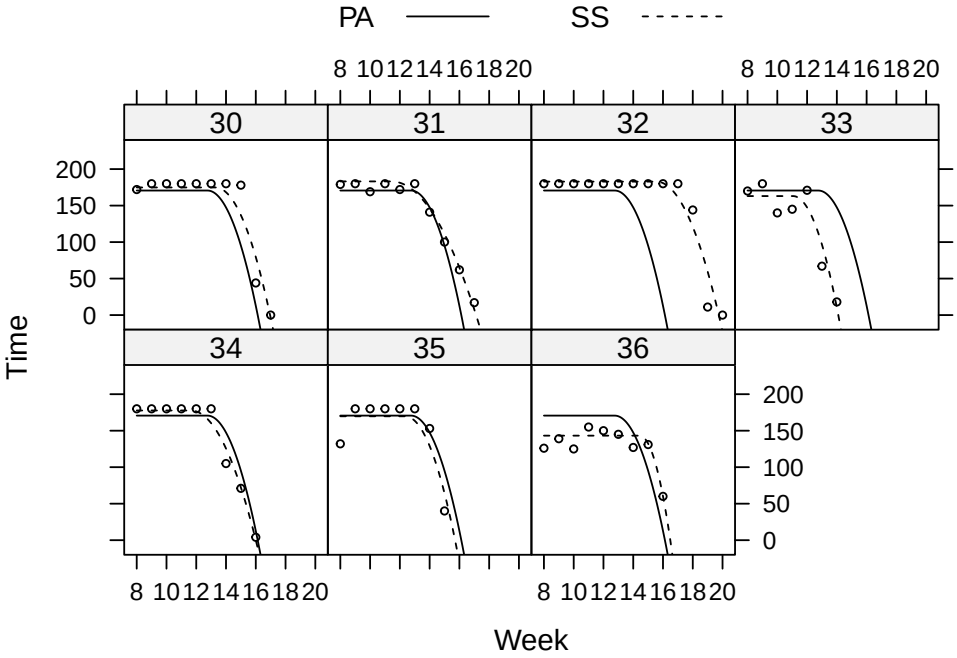
4) Calcular $y^{(r)}_{ijk} = y^{(r-1)}_{ijk} + \lambda^{(r-1)}_{ij}h_{ij}(\lambda^{(r-1)}_{ij})g_{ij}(\lambda^{(r-1)}_{ij})$.

5) Ajustar o modelo $y^{(r)}_{ijk} = \alpha_{ij} + \gamma_{ij}(t_k - \psi^{(r-1)})^2 I(t_k > \psi^{(r-1)}) + \lambda_{ij}h_{ij}(\lambda^{(r-1)}_{ij})g_{ij}(\lambda^{(r-1)}_{ij}) + e^{(r-1)}_{ijk}$ para obter $\alpha^{(r)}_{ij}$, $\gamma^{(r)}_{ij}$, $\lambda^{(r)}_{ij}$ e $\psi^{(r)} = [a_1 + a_2\exp(\lambda^{(r)}_{ij})]/[1 + \exp(\lambda^{(r)}_{ij})]$.

6) Parar se algum critério de convergência estiver satisfeito; em caso contrário, fazer $r = r + 1$ e repetir os passos 4-6.

Gráficos com os valores observados e previstos para as respostas de cada animal de um dos grupos analisados estão dispostos na Figura 2.

Figura 2: Gráficos de valores observados e previstos para as respostas



Comparações entre os parâmetros correspondentes às curvas esperadas associadas às combinações dos níveis dos diferentes tratamentos e sexos podem ser realizadas por meio de testes de Wald com base nos resultados do ajuste do modelo final. Os resultados são comparados com aqueles de modelos em que as curvas associadas aos períodos pré e pós ponto de mudança são retas, obtidos por meio das técnicas propostas por Muggeo et al. (2014) e Fasola et al. (2018).

A construção de algoritmos para o ajuste de modelos lineares mistos com mais do que um ponto de mudança é um tema interessante para futuras pesquisas.

AGRADECIMENTOS

Este trabalho recebeu apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, processo 3304126/2015-2) e Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processo 2013/21728-2), Brasil.

Referências

[1] Coatti, G.C. (2015). Avaliação do potencial terapêutico de pericitos e de células mesenquimais no camundongo SOD1, modelo animal para esclerose lateral amiotrófica. *Tese de doutorado, Departamento de Biocências, Universidade de São Paulo.* http://www.teses.usp.br/teses/disponiveis/41/41131/tde-14012016-143346/pt-br.php

[2] Fasola, S., Muggeo, V.M.R. and Küchenhoff, H. (2018). A heuristic, iterative algorithm for change-point detection in abrupt change models. *Compuational Statistics* 33, 997-1015.

[3] Muggeo, V.M.R., Atkins, D.C., Gallop, R.J. and Dimidjian, S. (2014). Segmented mixed models with random changepoints: a maximum likelihood approach with application to treatment for depression study. *Statistical Modelling* 14,: 293-313.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

POPULATION DYNAMICS EQUILIBRIUM AND EXTREME GROWTH

M. Fátima Brilhante[1], M. Ivette Gomes[2] e Dinis Pestana[3]

[1]Faculdade de Ciências e Tecnologia da Universidade dos Açores and Centro de Estatística e Aplicações da Universidade de Lisboa
[2]Centro de Estatística e Aplicações da Universidade de Lisboa and Instituto de Investigação Científica Bento da Rocha Cabral
[3]Centro de Estatística e Aplicações da Universidade de Lisboa and Instituto de Investigação Científica Bento da Rocha Cabral

ABSTRACT

Over the years the Verhulst model for population dynamics has been the building block for other population growth models. Since the solution of the Verhulst model and of some generalized versions are connected to max-geometric stable distributions or to generalized extreme value distributions, we prove that the exponent linked to the retroaction factor of some generalized models, whose solutions are not these distributions, determines on its own which limit law is appropriate for modeling extreme population growth.

Keywords and key sentences: Population Dynamics, Verhulst Model, Generalized Verhulst Models, Max-Geometric Stable Distributions, Generalized Extreme Value Distributions, Extreme Population Growth.

1. INTRODUCTION

Let $N(t)$ be the size of a population at time t. Under certain regularity conditions, Verhulst (1838) proposed the logistic differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}N(t) = rN(t)\left(1 - \frac{N(t)}{K}\right) \tag{1}$$

to model population dynamics, where $r > 0$ is the intrinsic growth rate and $K > 0$ is the carrying capacity, *i.e.* the limiting size the population may reach without disruptive effects on the availability of resources. In the right hand of equation (1), $N(t)$ is considered the growth factor and $1 - \frac{N(t)}{K}$ the retroaction factor, which is responsible for curbing down population growth. The solution of (1) is $N(t) = \frac{KN_0}{N_0+(K-N_0)\mathrm{e}^{-rt}}$, a member of the family of logistic functions, hence the name logistic (N_0 is the initial population size).
In spite of its limitations, the Verhulst model is still quite popular. One limitation of the model is being only suitable for modeling sustainable growth, or modeling stable populations. Therefore, over the years the Verhulst model has been used as basis for building several other

growth models, stating, for instance, that either $\frac{\mathrm{d}}{\mathrm{d}t}N(t)$ or $\frac{\mathrm{d}}{\mathrm{d}t}\ln N(t)$ is a decreasing function of the population density $\frac{N(t)}{K}$, such as in model (1). For example, the family of models

$$\frac{\mathrm{d}}{\mathrm{d}t}\ln N(t) = r\frac{1-\left(\frac{N(t)}{K}\right)^{\nu}}{\nu} \Leftrightarrow \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}N(t) = rN(t)\frac{\left(1-\left(\frac{N(t)}{K}\right)^{\nu}\right)}{\nu} & , \ \nu > 0 \\ \frac{\mathrm{d}}{\mathrm{d}t}N(t) = rN(t)\left(-\ln\left(\frac{N(t)}{K}\right)\right) & , \ \nu = 0 \end{cases}, \tag{2}$$

which is based on the Box-Cox family of transformations, contains the Verhulst model ($\nu = 1$). The subfamily for $\nu > 0$ was considered in Richards (1959), and the solution for $\nu = 0$ is $N(t) = K\exp\left(\ln\left(\frac{N_0}{K}\right)\mathrm{e}^{-rt}\right)$, which is known in population dynamics as the Gompertz growth model. This model is closely connected to the Gumbel distribution and has been used for modeling the growth of cancer tumors.
On the other hand, Blumberg (1968) extended the Verhulst equation by considering the hiperlogistic equation

$$\frac{\mathrm{d}}{\mathrm{d}t}N(t) = r\big(N(t)\big)^{\alpha}\left(1-\frac{N(t)}{K}\right)^{\beta}, \quad \alpha, \beta > 0\,. \tag{3}$$

However, equation (3) does not contain a closed form analytical solution, except for some special values of α and β. Generalizations of equation (3) and of family (2) for $\nu = 0$ was considered in Brilhante *et al.* (2011, 2012), in connection to the BetaBoop familiy of densities. We would like to point out that all extended versions of the Verhulst model are intended to be more flexible, allowing in some cases the possibility of modeling different types of unrestricted population growth. For more information on other growth models cf. Tsoularis (2001).
If we rewrite Verhulst's logistic equation solely as a function of the population density $\delta(t) = \frac{N(t)}{K}$, *i.e.* $\frac{\mathrm{d}}{\mathrm{d}t}\delta(t) = r\delta(t)\big(1-\delta(t)\big)$, the normalized solution $\delta(t) = \frac{1}{1+\mathrm{e}^{-rt}}$ belongs to the logistic family of distributions. Additionally, Brilhante *et al.* (2011) showed that the solution of the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}N(t) = rN(t)\left(-\ln N(t)\right)^{1+\xi}, \quad \xi \in \mathbb{R}\,, \tag{4}$$

which generalizes family (2) when $\nu = 0$ (and $K = 1$), belongs to the family of generalized extreme value (GEV) distributions for maxima. In particular, if $\xi > 0$, we have the Fréchet distribution, if $\xi = 0$, the Gumbel distribution and if $\xi < 0$, the Weibull distribution for maxima. But, in Statistics, more precisely, in Extreme Value Theory, the logistic distribution is a max-geometric stable distribution, *i.e.* a stable distribution for random maxima of sequences of independent and identically distributed (iid) random variables, with a geometric subordinator (cf. Rachev and Resnick, 1991), and GEV distributions are max-stable distributions of sequences of maxima of iid random variables.
As a note, max-stable distributions are necessarily of the generalized extreme value type

$$G_{\xi}(x) = \exp\Big(-(1+\xi x)^{-1/\xi}\Big), \quad 1+\xi x > 0,\ \xi \in \mathbb{R}\,,$$

whilst max-geometric stable distributions are of the type

$$H_{\xi}(x) = \frac{1}{1-\ln G_{\xi}(x)} = \frac{1}{1+(1+\xi x)^{-1/\xi}}, \quad 1+\xi x > 0,\ \xi \in \mathbb{R}\,.$$

The three types of max-geometric stable distributions are the log-logistic distribution ($\xi > 0$), the logistic distribution ($\xi = 0$) and the backward log-logistic distribution ($\xi < 0$).
Therefore, there seems to be a connection between population dynamics equilibrium and extreme growth for some extended Verhulst models. In the next section we shall investigate the type of connection that does occurs.

2. GENERALIZED VERHULST MODELS AND EXTREME GROWTH

In this section we begin by recalling some basic facts about max-stable distributions, which are used to prove the type of extreme limit population growth involved in each case presented here.

A distribution function (df) F is said to belong to the domain of attraction of a GEV distribution G_ξ for maxima if, and only if,

$$\lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)} = \begin{cases} \frac{x^\xi-1}{\xi} & , \xi \neq 0 \\ \ln x & , \xi = 0 \end{cases}, \quad x>0\,, \tag{5}$$

where $U(t) = F^{\leftarrow}\left(1-\frac{1}{t}\right)$, $t \geq 1$, is the reciprocal tail quantile function, $F^{\leftarrow}(y) = \inf\{x : F(x) \geq y\}$ is the generalized inverse function of F and $a(.)$ is a positive function.

Henceforth, we shall consider in the differential equations $K = 1$ in order to get a normalized solution $N(t) \in (0,1)$, more precisely, a df N. In fact, over time it has been observed that the form of the population curve has a sigmoid shape (cf. Smith, 1963).

Due to space restrictions, we shall mainly focus our attention on the Blumberg hiperlogistic equation (3). The solution satisfies the equation

$$\frac{(N(t))^{1-\alpha}}{1-\alpha}\, {}_2F_1(1-\alpha,\beta;2-\alpha;N(t)) = rt + C\,, \quad \alpha \notin \mathbb{N}\,, \tag{6}$$

where ${}_2F_1(a,b;c;z)$ is the hypergeometric function and C is a real constant. If $\alpha+\beta=2$, we get a closed form analytical solution for $N(t)$, namely

$$N(t) = \frac{1}{1+\left[\frac{t+C/r}{1/((1-\alpha)r)}\right]^{-1/(1-\alpha)}}\,,$$

since ${}_2F_1(a,b;b;z) = (1-z)^{-a}$. Observe that if $\alpha<1$ (or $\beta>1$), the solution belongs to the log-logistic family of distributions, and if $\alpha>1$ (or $\beta<1$), the solution belongs to the backward log-logistic family of distributions. Further observe that the Verhulst equation does satisfy the condition $\alpha+\beta=2$, since we have $\alpha=\beta=1$. Thus, a form of extreme stability becomes apparent for population growth in these particular cases.

On the other hand, from the close connection between max-geometric stable and max-stable distributions, it follows that if a df F is in the max-geometric domain of attraction of H_ξ, it will necessarily be in the max-domain of attraction of a GEV distribution G_ξ. Regardless of this connection, the max-geometric stable distributions have their own characterizations for domains of attraction.

If $\alpha+\beta \neq 2$ in (6), we can use the reciprocal tail quantile function, namely

$$U(t) = N^{\leftarrow}\left(1-\tfrac{1}{t}\right) = \frac{1}{r}\left[\frac{1}{1-\alpha}\left(1-\tfrac{1}{t}\right)^{1-\alpha}\, {}_2F_1\left(1-\alpha,\beta;2-\alpha;1-\tfrac{1}{t}\right) - C\right]\,,$$

to determine the limit distribution of N. Taking into account the properties of the hypergeometric function, we have $U(\infty) = \frac{{}_2F_1(1-\alpha,\beta;2-\alpha;1)}{(1-\alpha)r} < \infty$ if $\beta<1$ and $U(\infty)=\infty$ if $\beta \geq 1$. Using the results stated in the beginning of this section, we manage to prove that N is in the max-domain of attraction of a GEV distribution G_ξ, with $\xi = \beta - 1$.

If $\alpha = 1$ (and $\beta>0$) in equation (3), the reciprocal tail quantile function is now

$$U(t) = \frac{1}{r}\left[-\frac{1}{\beta}\left(\frac{t}{1-t}\right)^{\beta}\, {}_2F_1\left(\beta,\beta;\beta+1;\tfrac{t}{t-1}\right) - C\right]\,,$$

with $U(\infty) < \infty$ if $\beta < 1$ and $U(\infty) = \infty$ if $\beta \geq 1$. We also prove that N is in the max-domain of attraction of a GEV distribution G_ξ, with $\xi = \beta - 1$.
For $\alpha = m$ and $\beta = n$, where $m, n \in \mathbb{N}$, we have

$$U(t) = \frac{1}{r}\left[\sum_{k=2}^{m} \frac{a_k}{1-k} \frac{1}{\left(1-\frac{1}{t}\right)^{k-1}} + \sum_{j=2}^{n} \frac{b_j}{j-1} t^{j-1} + a_1 \ln\left(1-\tfrac{1}{t}\right) + b_1 \ln t - C\right],$$

where the a_k's and b_j's are positive constants. In this case $U(\infty) = \infty$, and it is quite straightforward to prove that the solution N belongs to the max-domain of a attraction of a GEV distribution G_ξ, with $\xi = n - 1$. When β is a non integer (and α an integer), we have not been able to prove that the solution N belongs to the max-domain of a attraction of a GEV distribution G_ξ, with $\xi = \beta - 1$. However, when we give values to α and β in (3) and solve the equation, all solutions so far confirm this result.
On the other hand, we get similar conclusions when considering the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} N(t) = r(N(t))^\alpha \left(-\ln N(t)\right)^\beta, \quad \alpha, \beta > 0, \tag{7}$$

a generalization of equation (4). In fact, equation (3) can also be considered a special case of (7), since $-\ln N(t) \approx 1 - N(t)$. However, the retroaction factor $-\ln N(t)$ in (4) is lighter than the retroaction factor $1 - N(t)$ in (1). Moreover, we already know that the max-stable distributions are the exact solutions if $\alpha = 1$ in (7), where $\xi = \beta - 1$. If $\alpha \neq 1$, the solution belongs to the max-domain of attraction of a GEV distribution G_ξ, with $\xi = \beta - 1$.
From the exposed above, it might seem, at first sight, a bit awkward that only β, the exponent linked to the retroaction factor, is important to establish which GEV distribution for maxima is at stake, and that α, the exponent of the growth factor, has no say in the matter of extreme population growth. Observe that since $1 - N(t) \in (0, 1)$, the retroaction factor $(1 - N(t))^\beta$ (or $\left(-\ln N(t)\right)^\beta$), will have a weaker control on population growth if $\beta > 1$, because $(1 - N(t))^\beta \to 0$ as $\beta \to \infty$, and that $(1 - N(t))^\beta$ will have a stronger control on population growth if $\beta < 1$, given that $(1 - N(t))^\beta \to \infty$ as $\beta \to 0$. The middle ground, or "equilibrium", is achieved whenever $\beta = 1$ ($\xi = 0$).

ACKNOWLEDGMENT

Funded by FCT-Fundação para a Ciência e Tecnologia, Portugal, Project UID/MAT/00006/2013.

References

[1] Blumberg, A.A. (1968). Logistic Growth Rate Functions. *Journal of Theoretical Biology*, 21, 42-44.

[2] Brilhante, M. F., Gomes, M.I., Pestana, D. (2011). BetaBoop Brings in Chaos. *CMSim - Chaotic Modeling and Simulation Journal*, 1, 39-50.

[3] Brilhante, M.F., Gomes, M.I., Pestana, D. (2012). Extensions of Verhulst Model in Population Dynamics and Extremes. *CMSim - Chaotic Modeling and Simulation Journal*, 4, 575-591.

[4] Rachev, S.T., Resnick, S. (1991) Max-Geometric Infinite Divisibility and Stability. *Comm. Statist. Stochastics Models*, 7, 191-218.

[5] Richards, F.J. (1959). A Flexible Growth Function for Empirical Use. *Journal of Experimental Botany*, 10(29), 290-300.

[6] Smith, F.E. (1963). Population Dynamics in Daphnia magna and a New Model for Population Growth. *Ecology*, 44(4), 651-663.

[7] Tsoularis, A. (2001). Analysis of Logistic Growth Models. *Res. Lett. Inf. Math. Sci.*, 2, 23-44.

[8] Verhulst, P.F. (1838). Notice sur la loi que la population poursuit dans son accroissement. *Corresp. Math. Physics*, 10, 113-121.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

MODELAGEM DE CAPTURAS EM PESO INFLACIONADAS DE ZEROS NO BAIXO RIO AMAZONAS

Júlio C. Pereira[1], Giovani L. Silva[2] e Victória J. Isaac[3]

[1]Universidade Federal de São Carlos - Brasil
[2]Dep. Matemática, Instituto Superior Técnico & CEAUL, Universidade de Lisboa
[3]Universidade Federal do Pará - Brasil

RESUMO

A análise de dados de captura e esforço gerados por pesca comercial são extremamente úteis na avaliação de estoques de pesca. Este trabalho foi motivado pela dificuldade em analisar dados de pesca no Baixo Rio Amazonas devido ao fenômeno de inflação de zeros capturas. O nosso objetivo é propor um modelo capaz de acomodar a inflação de zeros nas capturas em peso, permitindo uma melhor compreensão das variações da mesma relacionadas às variações no esforço e outras covariáveis disponíveis. Neste sentido, desenvolveu-se um modelo hierárquico bayesiano em três estágios, em que no primeiro estágio, modelou-se o número de viagens de pesca por local (N), de acordo com uma distribuição de Poisson. No segundo estágio, dado $N > 0$, definiu-se uma variável Bernoulli X, onde $X = 1$ indica ocorrência de capturas para uma determinada espécie, e $X = 0$, no caso contrário, no terceiro estágio, considerou-se o modelo probabilístico gama para o peso de pesca, denotado por Y, quando $N > 0$ e $X = 1$, onde o valor esperado de Y é proporcional ao número de visitas N. Esta abordagem fornece uma ferramenta útil para analisar a variação na captura por unidade de esforço como função de covariáveis quando os dados são inflacionados de zeros provenientes de ambas as fontes: abstinência da atividade pesqueira ($N = 0$) e ausência de captura na presença de atividade pesqueira ($X = 0$).

Palavras-chave: Inflação de zeros, Modelo Poisson composto, Estatística bayesiana, Pesca.

1. INTRODUÇÃO

A pesca, uma das atividades econômicas mais antigas do homem, possui enorme relevância em muitos países, sendo a sua produção uma das principais fontes de alimento na dieta de muitos povos. Por essa razão há uma crescente preocupação sobre a conservação dos recursos pesqueiros [2]. Cada vez mais organizações governamentais e comités internacionais procuram estabelecer políticas para a exploração sustentável de recursos pesqueiros com vista a assegurar a viabilidade a longo prazo deste setor económico e a garantir a produção de alimentos para as gerações futuras. A avaliação formal e criteriosa de um estoque de pesca passa pela construção de modelos quantitativos visando melhores predições, baseadas em

dados disponíveis. Além disso, a sua modelagem estatística permite entender como se dá a variação nas capturas em função do esforço e possíveis covariáveis associadas às embarcações e ao ambiente de pesca [1].
O presente trabalho foi motivado pela dificuldade encontrada por pesquisadores em modelar captura por unidade de esforço resultante de pescarias realizadas na região do Baixo Rio Amazonas. Essas pescarias são reportadas por cada viagem de pesca, identificadas com a latitude e longitude do centróide de uma área onde as pescarias ocorreram. Quando tomado o total mensal de capturas de uma espécie, associadas a cada centróide, tem-se o problema da inflação de zeros. Em muitas das localizações, tem-se captura mensal igual a zero para uma determinada espécie, enquanto que para outras localizações ocorrem capturas. Entretanto a inflação de zeros pode ocorrer por duas razões: i) não há atividade de pesca naquela localização, ou seja, nenhuma embarcação visitou aquele local em determinado mês; ii) há atividade de pesca i.e. houve uma ou mais visitas naquela localização, porém não houve captura da espécie em nenhuma das visitas.
O objetivo deste trabalho é desenvolver um modelo estatístico para acomodar adequadamente o excesso de zero nas capturas e assim possibilitar o melhor entendimento das variações da captura em peso, em função do esforço de pesca e de outras covariáveis associadas.

2. MOTIVAÇÃO

A motivação deste trabalho reporta-se aos dados de totais mensais de pescarias ocorridas no ano de 2004 na região do Baixo Amazonas e desembarcados no porto de Santarém - Brasil, bem como o número de viagens de pesca por localização, o total em peso capturado por espécie, o esforço de pesca e o tipo de ambiente de pesca: lagos (quando a pesca ocorre em áreas alagadas às margens dos rios) e rios (quando a pesca ocorre no canal do rio). No presente estudo, há interesse em modelagem das capturas da espécie Mapará (*Hypophthalmus marginatus and H. edentatus*) por ser a espécie que apresentou em geral o maior peso capturado. Mais detalhes dos dados podem ser obtidos em [6].

3. MODELO HIERÁRQUICO BAYESIANO

Para analisar estes dados de pesca propõe-se um modelo hierárquico bayesiano de três etapas. Na primeira etapa, descrevemos o número (N) de viagens de pesca por localização de acordo com uma distribuição de Poisson, i.e..

$$N_i \sim Poisson(\mu_i),\ i = 0, 1, 2, \ldots, n \tag{1}$$

em que n representa o número de unidades amostrais. O logaritmo da média μ_i foi considerado ser dependente de covariáveis disponíveis, isto é, $\log \mu_i = \beta_0 + \beta_1 river_i + \beta_2 mo_i$, em que $river$ e mo são variáveis *dummies* i.e. com valores iguais a zero ou um. Se $river = 1$, a pesca foi realizada no ambiente rio, enquanto $river = 0$ indica ambiente lago. A variável $mo = 1$ designa pesca realizada no período de março a outubro e $mo = 0$ no período de novembro a fevereiro. A fim de modelar possível superdispersão nas contagens, uma componente aleatória $v_i \sim N(0, \sigma^2)$ foi adicionada no $\log \mu_i$ da distribuição Poisson (1).
Na segundo etapa, dado $N > 0$, definiu-se uma variável de Bernoulli X com probabilidade q de sucesso (pesca da espécie), onde $X = 1$, se as capturas ocorreram para uma determinada espécie, e $X = 0$, se nada foi capturado para essa espécie, como mostrado a seguir.

$$X_i \mid N = n_i \sim Bern(q_i),\ \forall n_i > 0, \tag{2}$$

em que $P(X_i = 1 \mid N = n_i) = q_i$. O *logit* de q_i foi inicialmente considerado ser dependente das covariáveis disponíveis, $logit(q_i) = \mathbf{Z_i}\gamma$, sendo $\mathbf{Z_i}$ um vetor de covariáveis observadas e

associadas à probabilidade de sucesso. Em princípio, ajustou-se o modelo incluindo todas as covariáveis no *logit* de q, a fim de se selecionar posteriormente aquelas que eram significativas. Uma das versões dos modelos ajustados incluiu também o esforço de pesca como um *offset* na equação do preditor linear (sem resultados interessantes).
Finalmente, na terceira etapa modelou-se o peso denotado por Y, o qual tem valor zero quando $N = 0$ ou quando $N > 0$ e $X = 0$. Quando $N > 0$ e $X = 1$, modelou-se Y de acordo com uma distribuição gama, i.e.,

$$Y_i = \begin{cases} 0, & \text{se } N_i = 0, \\ 0, & \text{se } N_i > 0 \text{ e } X_i = 0, \\ \sum_{i=k}^{N_i} W_k, & \text{se } N_i > 0 \text{ e } X_i = 1, \end{cases} \tag{3}$$

em que W_k são pesos capturados para cada evento de pesca bem sucedido. As quantidades W_k são consideradas independentes e identicamente distribuídas de acordo com uma distribuição gama, i.e., $W_k \sim Gamma(a_0, b)$, logo o peso total capturado dado $N_i = n_i, n_i > 0$ and $X_i = 1$ é também distribuído de acordo com uma distribuição gama, $Y_i \mid N_i = n_i, X_i = 1 \sim Gamma(a_0 \times N_i, b)$ [4]. A média da distribuição gama foi também considerada depender de possíveis covariáveis.
O modelo hierárquico descrito nesta seção, bem como suas versões mais simples, foram ajustados seguindo uma abordagem bayesiana e implementados no software OpenBugs [5]. Posteriormente os modelos foram comparados usando-se critérios quer de bondade de ajuste quer de capacidade preditiva, isto é, *Deviance Information Criterion* (DIC) [7], *Watanabe-Akaike Information Criteria* (WAIC) [8], *log pointwise predictive density* Lppd e log conditional predictive ordinate (LCPO) [3].

4. RESULTADOS

A Tabela 1 apresenta os resultados dos critérios usados para a comparação dos modelos candidatos ajustados para o Mapará e a Tabela 2 mostra a proporção de predições corretas do número de viagens e do peso capturado. Considerando-se os resultados das Tabelas 1 e 2 o modelo $M4$ parece ser o mais indicado, pois apesar de não ser o melhor modelo de acordo com todos os critérios da Tabela 1 esse modelo foi o que apresentou a melhor capacidade preditiva conforme mostrado na Tabela 2.
As medianas das distribuições marginais a posteriori dos parâmetros do modelo $M4$ foram $\hat{\beta}_1 = -1.26$, $\hat{\beta}_2 = 0.43$, $\hat{\gamma_0} = -15.81$, $\hat{\gamma_2} = 7.6$, $\hat{m_0} = -3.26$, $\hat{m_2} = 3.58$, $\hat{a} = 0.61$, $\hat{\sigma_2} = 1.67$. De acordo com os valores obtidos para os parâmetros desse modelo, são esperadas as maiores médias do número de visitas no ambiente lago e durante o período de Março a Outubro. Os resultados também sugerem que o sucesso nas capturas depende do período do ano e as maiores probabilidades são esperadas entre Março e Outubro. Dado sucesso nas capturas o peso capturado também tende a ser maior nesse período. Além disso, a probabilidade de capturas e o peso médio capturado não dependem diretamente do ambiente de pesca.
Os resultados também mostraram que que o efeito aleatório na média do número de visitas N melhorou a capacidade preditiva do modelo, em particular melhorou a capacidade em predizer essa variável (N).

5. CONCLUSÕES

Essa abordagem forneceu uma ferramenta útil para se analisar a variação da captura por unidade de esforço como função de covariáveis quando os dados são inflacionados de zeros provenientes de duas fontes: ausência da atividade de pesca e ausência de capturas na presença de pesca.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pelos projeto FCT UID/MAT/00006/2013.

Tabela 1: Resultados dos criterios calculados para os modelos candidatos

	Model	*DIC*	*WAIC*	Lppd	LCPO
M1	λ : intercept + river + mo q : intercept + river + mo μ : intercept + mo + N	-1118.51	1048.31	-518.79	-524.18
M2	λ : intercept + river + mo q : intercept + mo + eff μ : intercept + mo + N	-1384.63	1050.16	-518.93	-525.27
M3	λ : intercept + river + mo q : intercept + mo + eff μ : intercept + mo + eff + N	1626.14	1637.33	-808.30	-818.77
M4	λ : river + mo + v q : intercept + mo + eff μ : intercept + mo + N	-1819.76	1078.09	-489.27	-541.00

Tabela 2: Proporção de predições corretas

Model	$N = 0$	$N > 0$	$Y = 0 \mid N > 0$	$Y > 0 \mid N > 0$
M1	0.26	0.89	0.80	0.28
M2	0.29	0.91	0.93	0.66
M3	0.29	0.89	0.93	0.66
M4	1	1	0.91	0.68

Referências

[1] FAO (2006). Stock assessment for fishery management: a framework guide to the stock assessment tools of the fisheries management science programme (FMSP). *Fisheries Technical Paper* 487.

[2] Faveret Filho, P., Siqueira, S.H. (1997). Panorama da pesca marítima no Mundo e no Brasil. *BNDES Setorial* 5, 185–198.

[3] Gelman, A., Hwang, J., Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* 24, 997–1016.

[4] Hogg, R.V., McKean, J.W., Craig A.T. (2004). *Introduction to Mathematical Statistics* (6th ed.). Prentice Hall, Upper Saddle River, New Jersey.

[5] Lunn, D., Spiegelhalte, D., Thomas A., Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine* 28 (25), 3049–3067.

[6] Pinaya, W.H.D., Lobon-Cervia, F.J., Pita, P., Buss de Souza, R., Freire, J., Isaac, V.J. (2016). Multispecies fisheries in the Lower Amazon River and its relationship with the regional and global climate variability. *PLoS ONE* 11(6), e0157050.

[7] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* 64, 583–639.

[8] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11, 3571–3594.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

MEDIDAS DE FIABILIDADE DE CLASSIFICAÇÃO BINÁRIA COM BASE NUMA VARIÁVEL QUANTITATIVA – UMA COMPARAÇÃO VIA SIMULAÇÃO

Rui Santos[1], Miguel Felgueiras[2], João Paulo Martins[3] e Liliana Ferreira[4]

[1]Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, CEAUL – Centro de Estatística e Aplicações, rui.santos@ipleiria.pt
[2]Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, CEAUL – Centro de Estatística e Aplicações, Centre of Applied Research in Management and Economics, mfelg@ipleiria.pt
[3]Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, CEAUL – Centro de Estatística e Aplicações, jpmartins@ipleiria.pt
[4]Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria, Centro de Matemática, Aplicações Fundamentais e Investigação Operacional, liliana.ferreira@ipleiria.pt

RESUMO

Sob diversos cenários de classificação binária gerados via simulação utilizando várias distribuições para a caracterização das duas subpopulações, são comparados os valores obtidos em medidas de fiabilidade de classificação binária com o objetivo de aferir a sua adequação sob diferentes condições. Em particular, são comparados os valores da área sob a curva ROC, integral e parcial utilizando distintas amplitudes, com o desempenho (sensibilidade e especificidade) no ponto de corte "ótimo", correspondente ao valor máximo do índice de Youden, à distância mínima ao ponto ideal e à probabilidade máxima de concordância na classificação.

Palavras chave: Área integral sob a curva ROC, Área parcial sob a curva ROC, Classificação binária, Especificidade, Sensibilidade, Simulação.

1. INTRODUÇÃO

Numa população com N indivíduos, considere-se uma infeção com taxa de prevalência p. Seja $X_i \sim \text{Ber}(p)$, $i = 1, \cdots, N$, uma variável aleatória (v.a.) que representa a presença ($X_i = 1$) ou a ausência ($X_i = 0$) da infeção no indivíduo i; e seja Y_i a quantidade da substância de interesse no i-ésimo indivíduo, caracterizada pela distribuição $\text{D}_0(\theta_0)$ se $X_i = 0$ e $\text{D}_1(\theta_1)$ se $X_i = 1$. Neste contexto, seja t o ponto de corte da classificação binária (infetado *versus* saudável) baseada na observação do valor da v.a. Y_i. Assim sendo, temos definido o seguinte sistema de classificação (poderiam ser utilizadas as desigualdades opostas, mas o raciocínio seria análogo):

- $Y_i \leq t \Rightarrow X_i^-$ (resultado negativo, indivíduo classificado como saudável);

- $Y_i > t \Rightarrow X_i^+$ (resultado positivo, indivíduo classificado como infetado).

Deste modo, para cada ponto de corte t, podemos determinar o valor das medidas de fiabilidade da classificação, nomeadamente a especificidade (ou fração de verdadeiros negativos) que corresponde à probabilidade de se obter um resultado negativo num indivíduo saudável, i.e.

$$\varphi_{_e} = \mathrm{P}\left(X_i^-|X_i = 0\right) = \mathrm{P}\left(Y_i \leq t|X_i = 0\right) = F_{_{\mathrm{D}_0}}(t),$$

e a sensibilidade (fração de verdadeiros positivos) que corresponde à probabilidade de se obter um resultado positivo num indivíduo infetado, i.e.

$$\varphi_{_s} = \mathrm{P}\left(X_i^+|X_i = 1\right) = \mathrm{P}\left(Y_i > t|X_i = 1\right) = 1 - F_{_{\mathrm{D}_1}}(t) = \overline{F}_{_{\mathrm{D}_1}}(t).$$

2. A ÁREA SOB A CURVA ROC E O ÍNDICE ϕ

A curva ROC (*Receiver Operating Characteristic*) permite visualizar a evolução da $\varphi_{_s}$ e da $\varphi_{_e}$ quando percorremos todos os possíveis valores t para o ponto de corte, revelando todos os pares $(1 - \varphi_{_e}, \varphi_{_s})$ que podem ser representados por $(x, \mathrm{ROC}(x))$. Por este motivo, a curva ROC é utilizada para determinar o ponto ótimo de uma metodologia de classificação binária, assim como para comparar o desempenho de diferentes metodologias [**?**, **?**, **?**, **?**, **?**].
A área sob a curva ROC (AUC) representa o valor médio de $\varphi_{_s}$ para todos os valores de $\varphi_{_e}$, mas também pode ser interpretada como a probabilidade de corretamente classificar um par (onde 0.5 significa ausência de fiabilidade, como numa classificação aleatória, e 1 corresponde a uma classificação perfeita). O valor da área está igualmente relacionado com a estatística de teste de Wilcoxon-Mann-Whitney, permitindo efetuar inferência sobre a curva ROC. A AUC é, possivelmente, a medida mais comum para aferir o desempenho de uma metodologia de classificação binária [**?**, **?**, **?**]. Todavia, esta medida tem em consideração todos os possíveis valores do ponto de corte, mesmo aqueles que sejam desadequados na prática por terem associado um valor demasiado reduzido da $\varphi_{_s}$ ou da $\varphi_{_e}$. Assim, para avaliar unicamente o desempenho nos valores do ponto de corte para os quais a metodologia apresenta um desempenho satisfatório, pode ser utilizada a área parcial sob a curva ROC (pAUC) [**?**, **?**, **?**, **?**, **?**] determinada através de

$$\mathrm{pAUC}(x_0, x_1) = \int_{x_0}^{x_1} \mathrm{ROC}(x)\ \mathrm{d}x,$$

verificando $\mathrm{pAUC}(0, 1) = \mathrm{AUC}$ e $\frac{1}{2}\left(x_1^2 - x_0^2\right) \leq \mathrm{pAUC}(x_0, x_1) \leq x_1 - x_0$. Para a interpretar, de modo análogo à AUC, a pAUC pode ser estandardizada através de

$$\mathrm{spAUC}(x_0, x_1) = \frac{1}{2}\left(1 + \frac{\mathrm{pAUC}(x_0, x_1) - \frac{1}{2}\left(x_1^2 - x_0^2\right)}{x_1 - x_0 - \frac{1}{2}\left(x_1^2 - x_0^2\right)}\right).$$

Todavia, a utilização de pAUC ou spAUC requer a definição do intervalo (x_0, x_1) a partir dos valores de interesse, usualmente os valores mais elevados da $\varphi_{_e}$ ou da $\varphi_{_s}$.
O índice ϕ é outra medida de fiabilidade [**?**, **?**], correspondendo à probabilidade ϕ que verifica (para determinado valor t) $\varphi_{_e} = \varphi_{_s} = \phi$ sendo que, nos casos em que este valor não exista (e.g. em distribuições discretas), a distância entre $\varphi_{_s}$ e $\varphi_{_e}$ deve ser minimizada e $\phi = \frac{1}{2}(\varphi_{_s} + \varphi_{_e})$.

3. O PONTO DE CORTE "ÓTIMO"

Na prática, muitas vezes utilizamos unicamente um ponto de corte, pelo que o conhecimento do desempenho nesse ponto poderá ser suficiente. A determinação do valor do ponto de corte

"ótimo" é uma decisão que depende de diversos fatores, tais como a gravidade da infeção, o risco da infeção se não for diagnosticada ou os efeitos colaterais do tratamento. Deste modo, pode ser relevante decidir entre ter uma sensibilidade elevada ou uma especificidade elevada. No entanto, na ausência de fatores clínicos que conduzam a privilegiar uma destas medidas, o ponto de corte "ótimo" pode ser determinado através da otimização de critérios, tais como a maximização do índice de Youden [**?**, **?**, **?**, **?**], definido por

$$\mathrm{YI} = \varphi_{e} + \varphi_{s} - 1 = F_{\mathrm{D}_0}(t) - F_{\mathrm{D}_1}(t),$$

ou a minimização da distância ao ponto ideal (com $\varphi_{e} = \varphi_{s} = 1$) [**?**, **?**] que corresponde a maximizar

$$\mathrm{DI} = 1 - \sqrt{(1-\varphi_{e})^2 + (1-\varphi_{s})^2} = 1 - \sqrt{\overline{F}^2_{\mathrm{D}_0}(t) + F^2_{\mathrm{D}_1}(t)};$$

ou a maximização da probabilidade de concordância na classificação dicotómica [**?**] dada por

$$\mathrm{CP} = \varphi_{e}\varphi_{s} = F_{\mathrm{D}_0}(t)\overline{F}_{\mathrm{D}_1}(t).$$

4. SIMULAÇÃO

As simulações realizadas, através do *software* R (recorrendo aos *packages* ROCR e pROC) utilizando 10^3 réplicas, têm como objetivo analisar várias medidas de fiabilidade da classificação [AUC, ϕ e spAUC nos intervalos $(0.9, 1)$, $(0.75, 1)$, $(0.5, 1)$ e $(\phi - 0.05, \min\{\phi + 0.05, 1\})$ para a especificidade e para a sensibilidade], bem como diversos índices de determinação do ponto de corte "ótimo" (YI, DI e CP). Foram estudados diversos cenários, utilizando distintas dimensões da amostra $n_0 = n_1 \in \{50, 100, 250, 500, 1000\}$ e distribuições para a caracterização das duas subpopulações, supondo $\mathrm{D}_0 = \mathrm{D}_1$ mas com $\theta_0 \neq \theta_1$, nomeadamente:

- Normal: $\mathrm{N}(\mu, \sigma)$, $\mu_0 = 0$, $\sigma_0 = 1$, $\mu_1 = 2$ e $\sigma_1 \in \{2/3, 1, 1.5, 2, 3\}$;
- Gama: $\Gamma(\alpha, \beta)$, $\alpha_0 = 2$, $\beta_0 = 1$, $\alpha_1 \in \{6, 9, 12\}$ e $\beta_1 \in \{1, 3\}$,
- Binomial: $\mathrm{B}(n = 20, p)$, $p_0 = 0.25$ e $p_1 \in \{0.3, 0.4, 0.5\}$;
- Geométrica: $\mathrm{G}(p)$, $p_0 = 0.2$ e $p_1 \in \{0.1, 0.02\}$.

5. CONCLUSÕES

Dando continuidade a análises previamente apresentadas [**?**, **?**], os resultados obtidos nas simulações realizadas evidenciam que, na maioria das situações, AUC, spAUC e ϕ estão fortemente correlacionados e, portanto, parecem avaliar os mesmos critérios de fiabilidade. Ainda assim, salientam-se as seguintes conclusões:

- AUC mostra menos variabilidade que spAUC, nomeadamente em amostras pequenas e em situações com menor fiabilidade;
- spAUC determinada no intervalo $(\phi - 0.05, \min\{\phi + 0.05, 1\})$ mostra menor variabilidade que nos intervalos $(0.9, 1)$, $(0.75, 1)$ ou $(0.5, 1)$;
- spAUC parece fornecer melhores resultados quando determinado sobre a especificidade (comparando com a sensibilidade), mas nem sempre;
- ϕ parece, em muitos casos, ter maior correlação com YI, DI, CP que AUC ou spAUC;
- os pontos de corte definidos por YI, DI e CP podem dar origem a fiabilidades significativamente distintas, razão pela qual é aconselhável comparar os seus desempenhos em cada aplicação.

AGRADECIMENTOS

Este trabalho foi financiado por Fundos Nacionais através da FCT — Fundação para a Ciência e a Tecnologia, no âmbito dos projetos UID/MAT/00006/2013 e UID/MAT/04561/2013.

Referências

[1] Dodd, L.E. Pepe, M.S. (2003). Partial AUC estimation and regression. *Biometrics* 59, 614–623.

[2] Fluss, R., Faraggi, D., Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point, *Biom J* 47, 458–472.

[3] Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med* 4, 627–635.

[4] Hanley, J.A., McNeil, J.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36.

[5] Jiang, Y., Metz, C.E., Nishikawa, R.M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* 201, 745–750.

[6] Krazanowski, W.J., Hand, D.J. (2009). *ROC Curves for Continuous Data*. CRC press, New York.

[7] Liu, X. (2012). Classification accuracy and cut point selection. *Stat Med* 31, 2676–2686.

[8] Ma, H., Bandos, A., Rockette, H., Gur, D. (2013). On use of partial area under the ROC curve for evaluation of diagnostic performance. *Stat Med* 32, 3449–3458.

[9] Ma, H., Bandos, A., Gur, D. (2015). On the use of partial area under the ROC curve for comparison of two diagnostic tests. *Biom J* 57, 304–320.

[10] Metz, C.E. (2008). ROC analysis in medical imaging: a tutorial review of the literature. *Radiol Phys Technol* 1, 2–12.

[11] Pepe, M.S. (2003). *Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford.

[12] Perkins, N.J., Schisterman, E.F. (2006). The inconsistency of "optimal" cut-points using two ROC based criteria. *Am J Epidemiol* 163, 670–675.

[13] Powers, (2011). Evaluation: From Precision, Recall and F-Score to ROC, Informedness, Markedness & Correlation. *J Mach Learn Tech* 2, 37–63.

[14] Santos, R., Martins, J.P., Felgueiras, M. (2015). An Overview of Quantitative Continuous Compound Tests. In Bourguignon, J.P., Jeltsch, R., Pinto, A., Viana, M. (Eds.): *Dynamics, Games and Science, CIM Series in Mathematical Sciences* 1, 627–641.

[15] Santos, R., Felgueiras, M., Martins, J.P. (2015). Discrete Compound Tests and Dorfman's Methodology in the Presence of Misclassification. In Kitsos, C.P. *et al.* (Eds.): *Theory and Practice of Risk Assessment, Springer Proceedings in Mathematics & Statistics* 136, 85–98.

[16] Santos, R., Martins, J., Felgueiras, M., and Ferreira, L. (2017). Binary Classification Based on a Quantitative Variable – an Accuracy Comparison by Simulation. *Proceedings of 17th CMMSE*, 1883–1886.

[17] Santos, R., Felgueiras, M., Martins, J.P. e Ferreira, L. (2017). Área integral versus área parcial sob a curva ROC, *Programa e Resumos do XXIII Congresso da Sociedade Portuguesa de Estatística*, 238–239.

[18] Walter, S.D. (2005). The partial area under the summary ROC curve. *Stat Med* 24, 2025–2040.

[19] Witten, E. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med* 4, 627–635.

[20] Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3, 32–35.

[21] Zhou, X.H., Obuchowski, N.A., McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*. Wiley & Sons, New York.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

IMPUTAÇÃO MÚLTIPLA BASEADA NO ALGORITMO MONTE CARLO VIA CADEIA DE MARKOV (MCMC) PARA A ESTIMAÇÃO DE PARÂMETROS GENÉTICOS QUANTITATIVOS E SELEÇÃO DE GENÓTIPOS

Maria Márcia Pereira Sartori [1], Lucas Vasconcelos Vieira[1], Gabriela Nunes da Piedade[1]; Maurício Dutra Zanotto[1]

[1] Departamento de Produção e Melhoramento Vegetal, FCA/UNESP/Botucatu, São Paulo, Brasil.
contato: mmpsartori@fca.unesp.br

RESUMO

Na condução de experimentos agrícolas, pesquisadores frequentemente deparam-se com a perda de observações. Como solução para este problema, existem diversos métodos descritos na literatura para a imputação de dados. No entanto, apesar da facilidade com a qual podem ser executadas em muitos softwares, as implicações de tais métodos são muitas vezes confusas, além disso diferentes métodos podem resultadar em diferentes conclusões para o mesmo problema. Nesse sentido, o objetivo desse trabalho foi avaliar os efeitos de três métodos de imputação de dados sobre a predição de parâmetros genéticos quantitativos e sobre o ordenamento de genótipos. Dados de um experimento de um ensaio com Mamona (*Ricinus communis* L.) realizado na Universidade Estadual Paulista (UNESP/Botucatu) foram submetidos a um estudo de simulação implementada utilizando a linguagem IML (*interatictive matrix language*) e os procedimentos MI e MIXED no programa estatístico SAS 9.4. Para isso, 10 % dos dados foram retirados aleatoriamente da matriz inicial de dados e posteriormente submetidos a imputação pela média, a imputação simples estocástica e a imputação múltipla baseada no algoritmo Monte Carlo via Cadeia de Markov. Não houve diferença entre os métodos de imputação avaliados para a estimação dos componentes de variância e para o cálculo de herdabilidade. Contudo, a seleção de linhagens foi afetada dependendo do método de imputação utilizado.

Palavras e frases chave: imputação simples; imputação múltipla; MCMC.

1. INTRODUÇÃO

Frequentemente, os pesquisadores precisam lidar com algum grau de desbalanceamento no conjunto de dados. Embora o desequilíbrio possa muitas vezes ser uma característica inerente dos projetos experimentais, projetos inicialmente balanceados podem ficar desbalanceados quando uma ou mais observações são perdidas por circunstâncias imprevistas [1]. Dados discrepantes ou digitados erroneamente, falta de material, perda de plantas e parcelas devido a fatores biológicos ou ambientais, grupos experimentais com diferentes números de repetições por tratamento e a não

avaliação de todas as combinações genótipo-ambiente são exemplos que podem gerar conjuntos de dados com delineamentos desbalanceados ou incompletos (*missing data*). Como consequência, as análises estatísticas podem resultar em conclusões pouco verdadeiras dos experimentos e, portanto, afetam por exemplo a estimação de parâmetros genéticos e de valores genotípicos em programas de melhoramento [2].

Diversos métodos têm sido descritos na literatura para lidar com a perda de observações, sendo a imputação de dados preferencialmente escolhida por muitos pesquisadores [3]. No entanto, as implicações de tais métodos sobre as conclusões finais dos experimentos podem carretar em discrepâncias nos resultados dependendo do método escolhido. As técnicas de imputação simples, dentre as quais destaca-se a imputação pela média, são bastante conhecidos devido a facilidade com a qual podem ser implementados. Contudo, tais métodos apresentam diversas desvantagens pois não consideram a incerteza do valor correto a ser imputado. No caso da imputação pela média, as variâncias e covariâncias podem ser subestimadas além do comprometimento da correlação entre variáveis e dos intervalos de confiança. Outro método bastante empregado é o da imputação baseada em regressão linear, que pode melhorar a estimação dos valores a serem imputados, principalmente devido a não atenuação dos coeficientes de correlação e a inclusão de covariáveis no ajuste do modelo. Todavia, esse método pode inflacionar o poder de predição do modelo e aumentar a multicolinearidade.

Uma alternativa para corrigir tais problemas é a imputação simples estocástica, que substitui o valor ausente por um valor predito por imputação de regressão acrescido de um residual baseado em certa distribuição de probabilidade. Contudo, os métodos de imputação simples, mesmo a estocástica, podem produzir estimativas viesadas dos parâmetros populacionais, por não considerar a incerteza entre imputações. Isso acontece em decorrência de apenas um valor ser estimado para cada observação ausente, levando a subestimação dos erros padrões, ao encurtamento dos intervalos de confiança e consequentemente a um aumento do erro estatístico do tipo I. Por isso, nos últimos anos, tem-se dado preferência por métodos de imputação múltipla. Tais método permitem que mais de um valor seja considerado para cada observação ausente. Assim, a incerteza sobre o valor correto a ser imputado é considerado no modelo de imputação, aumentando assim o erro padrão e diminindo o viesamento. O método MCMC é amplamente utilizado para a inferência bayesiana e é o algoritmo iterativo mais popular para imputações múltiplas [4]. O algoritmo MCMC objetiva simular distribuições multivariadas, que tenham como limite uma cadeia de Markov estacionária.

Nesse sentido, o objetivo desse trabalho foi avaliar o efeito da imputação simples estocástica e da imputação múltipla baseada no método MCMC sobre a variância genética das linhagens avaliadas, o cálculo da herdabilidade e a seleção das linhagens.

2. MATERIAIS E MÉTODOS

Os dados utilizados nesse estudo foram provenientes de ensaios com Mamona (*Ricinus communis* L.) do Departamento de Produção e Melhoramento Vegetal da Faculdade de Ciências Agronômicas de Botucatu (FCA/UNESP). No ensaio, 12 híbridos foram avaliados em 2 ambientes (Lins e Penápolis) e 2 anos (2008 e 2009) utilizando um delineamento em blocos para constatação dos caracteres agronomicos de produtividade de grãos, teor de óleo, produtividade de óleo e altura de plantas.

A partir do conjunto original de dados referente a variável teor de óleo, os dados foram submetidos a um estudo de simulação implementada utilizando a linguagem IML (*interatictive matrix*

language) para a retirada aleatória de 10% dos dados gerando um padrão arbitrário de ausência. Posteriormente, os conjuntos de dados com observações ausentes foram submetidos a imputação pela média, a imputação simples estocástica e a múltipla via algoritmo MCMC utilizando o procedimento MI no software SAS 9.4, com as opções FCS, NBITER=1 e NIMPUTE=1 para a imputação simples estocástica, e com as opões MCMC para a imputação múltipla. Em seguida, os parâmetros genéticos e valores genótipos foram obtidos utilizando o procedimento MIXED, no qual o efeito de genótipo foi considerado aleatório a fim da obtenção dos valores genótipos através dos melhores preditores não-enviesados (*BLUP*), enquanto que a estimação dos componentes de variância foi realizada via máxima verossimilhança restrita (*REML*) conforme recomendado por [5].

2. RESULTADOS E DISCUSSÕES

Não houve diferença entre os métodos de imputação avaliados para a estimação dos componentes de variância, e consequentemente, para o cálculo de herdabilidade considerando uma porcentagem de 10% de ausência na matriz de dados (Tabela 1). Contudo, houve diferença no ordenamento dos genótipos para Lins e Penápolis no ano de 2008, sendo que em Penápolis as diferenças foram maiores (Tabela 2 e 3).

Locais	Anos	CO	ISM	ISE	MCMC
Lins	2008	0.79	0.79	0.76	0.76
	2009	0.95	0.95	0.93	0.95
Penápolis	2008	0.42	0.42	0.37	0.42
	2009	0.95	0.95	0.90	0.93

Tabela 1. Estimativa da herdabilidade no sentido amplo (h^2) a partir do conjunto original dos dados (CO), do método de imputação simples pela média (ISM), da imputação simples estocástica (ISE) e da imputação baseada no método Monte Carlo via Cadeia de Markov (MCMC).

CO			ISM			ISE			MCMC	
Gid	Estimativa		Gid	Estimativa		Gid	Estimativa		Gid	Estimativa
3	-2.867		3	-2.867		3	-2.732		3	-2.7216
12	-1.9093		12	-1.9093		12	-1.8037		12	-1.7974
9	-1.4823		9	-1.4823		9	-1.3898		9	-1.3853
1	-1.3986		1	-1.3986		1	-1.3086		1	-1.3045
10	-1.0854		10	-1.0854		10	-1.005		10	-1.0022
4	0.3225		4	0.3225		4	0.3597		4	0.3565
8	0.4641		8	0.4641		8	0.497		8	0.4932
6	0.644		6	0.644		6	0.6714		6	0.6668
2	0.9963		2	0.9963		2	1.0129		2	1.0068
5	**2.0173**		**5**	**2.0173**		**11**	**1.742**		**11**	**1.6168**
7	**2.0997**		**7**	**2.0997**		**7**	**1.9533**		**5**	**1.9921**
11	**2.1989**		**11**	**2.1989**		**5**	**2.0026**		**7**	**2.0788**

Tabela 2. Ordenamento de genótipos (Gid) utilizando a método de BLUP a partir do conjunto original dos dados (CO), a imputação simples pela média (ISM), imputação simples estocástica (ISE) e imputação baseada no método Monte Carlo via Cadeia de Markov (MCMC) para Lins 2008.

CO			ISM			ISE			MCMC	
Gid	Estimativa		Gid	Estimativa		Gid	Estimativa		Gid	Estimativa
10	-1.8254		10	-1.8254		10	-1.5369		10	-1.7929
1	-1.1276		1	-1.1276		9	-0.793		9	-1.0964
9	-0.6153		9	-0.6153		4	-0.6852		1	-0.6588
3	-0.09314		3	-0.09314		1	-0.646		2	-0.04767
2	-0.01912		2	-0.01912		3	-0.02022		3	-0.04237
6	0.1343		6	0.1343		11	0.2364		4	0.04202
4	0.3141		4	0.3141		12	0.2701		12	0.1021
12	0.3819		12	0.3819		2	0.3264		11	0.4898
5	0.5659		5	0.5659		5	0.5568		8	0.5531
11	0.6567		11	0.6567		8	0.5744		5	0.6236
8	0.6778		8	0.6778		6	0.8243		6	0.8159
7	0.9499		7	0.9499		7	0.893		7	1.0117

Tabela 3. Ordenamento de genótipos (Gid) utilizando a método de BLUP a partir do conjunto original dos dados (CO), da imputação simples pela média (ISM), da imputação simples estocástica (ISE) e da imputação baseada no método Monte Carlo via Cadeia de Markov (MCMC) para Penápolis 2008.

3. CONCLUSÕES

O sucesso dos programas de melhoramento genético de plantas está intrinsicamente relacionado com a precisão das estimativas e predições dos parâmetros populacionais dos genótipos sob avaliação. Apesar dos resultados mostrarem diferenças mínimas para o cálculo de herdabilidade, o ranqueamento das linhagens foi afetado pelo método de imputação utilizado. Portanto, é necassário cautela no uso de tais métodos.

AGRADECIMENTOS

Ao Conselho Nacional de Pesquisa (CNPq) e a Facudade de Ciências Agrárias (FCA) da Universidade Estadual Paulista “Júlio de Mesquita Filho (UNESP)

Referências

[1] Hocking, R.R. 2005. Mixed Models III: Unbalanced Data. p. 539–587. In Methods and Applications of Linear Models. John Wiley & Sons, Inc.

[2] Piepho, H.P., J. Möhring, A.E. Melchinger, and A. Büchse. 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161(1–2): 209–228.

[3] Das, S., A.K. Paul, S.D. Wahi, and U.K. Pradhan. 2017. Comparative Performance of Imputation Methods for Different Proportions of Missing Data in Classification of Crop Genotypes. (April).

[4] Little, R.J.A., D.B. Rubin, R.J.A. Little, and D.B. Rubin. 2002. 10. Bayes and Multiple Imputation. Stat. Anal. with Missing Data: 200–220Available at http://dx.doi.org/10.1002/9781119013563.ch10.

[5] Resende, M.D.V. de. 2004. Métodos Estatísticos Ótimos na Análise de Experimentos de Campo.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

ESTIMATION OF REFERENCE EQUATIONS FOR SPIROMETRY FOR NON-CAUCASION POPULATION

Carina Silva[1,2], Anália Matos[1] e Tânia Duarte[1,3]

[1]Escola Superior de Tecnologia da Saúde de Lisboa, IPL
[2]Centro de Estatística e Aplicações, Universidade de Lisboa (CEAUL)
[3]Instituto Clínico de Alergologia, Lisboa.

ABSTRACT

Spirometry is the single most important test for the evaluation of respiratory function. The interpretation is based on the comparison of the measured data with the predicted values obtained from a reference population. The American Thoracic Society (ATS)/European Respiratory Society (ERS) recommend ethnicity-specific reference standards in 2005. The Global Lung Initiative task force announced spirometric reference equations (GLI 2012) for multi-ethnic populations. However, in most lung function laboratories, the choice of reference values are the recommendations from European Community of Steel and Coal (ECSC), commonly used in Europe. Reference equations derived from spirometry data locally collected in a practical setting by well-trained personnel might be more appropriate for everyday use than generally used equations based on data from scientific studies in the distant past.

The aim of this study was to estimate spirometric reference equations for the non-caucasian population and compare the results to those from the Global Lung Initiative European Community (GLI 2012) and from Steel and Coal (ECSC).

It was used the Generalized Additive Models for Location and Shape (GAMLSS), used also in GLI 2012, to estimate equations for the spirometric parameters considering age and height both for women and man.

Keywords and key sentences: Spirometry, GLI2012, GAMLSS .

1. INTRODUCTION

Spirometry is the single most important test for the evaluation of respiratory function, considered an essential tool in the diagnosis and evaluation of individuals with respiratory pathology [**?**]. Their result supports the clinical decision, so its correct interpretation and classification are fundamental. The interpretation is based on the comparison of the measured data with the predicted values obtained from a reference population [**?**]. The predicted values, vary with age, sex, standing height, and ethnic group, and are obtained using reference equations[**?**]. Respiratory impairment is influenced by ethnic differences[**?**]. For the interpretation of spirometry it is necessary to identify the correct approach to adjust spirometry reference values for

ethnicity, which often is poorly defined and inconsistently applied. There has yet to be a workable definition of what constitutes ethnicity and the selection of which one applies to a given individual has significant implications for the interpretation of their test results and is often arbitrary. Frequently, a correction factor is used if the subject is of non-Caucasian descent[**?**]. While this approach may be appropriate for some people, the inherent variability of the population based on age, height and gender is not taken into account, therefore the use of a fixed percentage reduction is unlikely to be valid in all patients[**?**]. The variation of ethnic spirometric reference equations has been previously reported in the literature. The guidelines, published in 2005 by the American Thoracic Society (ATS)/European Respiratory Society (ERS), recommend ethnicity-specific reference standards. The Global Lung Initiative task force announced spirometric reference equations (GLI 2012) derived from data collected from healthy nonsmokers in the age group of 3–95 years from 33 countries. The GLI 2012 equations provided multi-ethnic values and the lower limit of normal (LLN) for spirometry. However, in most lung function laboratories, the choice of reference values are the recommendations from European Community of Steel and Coal (ECSC), commonly used in Europe. As each recommendation uses a different population, it is clear that the obtained reference equations must also be different, creating potential problems for the clinicians and technologists in the interpretation of the spirometric results. Thus, it is important to identify the implications of adopting different equations.

The participants data such as nationality, workplace, smoking history, age, height, and weight were collected using a questionnaire before the spirometry ,to eliminate subjects who not fulfill the inclusion criteria. The examinations were performed by a qualified and certified technician with sufficient training to ensure that proper testing procedures were followed. Individuals from 18 to 75 years old, in an Black population and without clinical history of pulmonary changes, submitted the form application and spirometry realization. The spirometry tests were performed between January 2018 and March 2018, and the sample was collected from pharmacies of the group HOLON placed in several cities of Portugal and in a Cape Verdean association resulting in a total of 204 participants. After excluding the participants concerning the exclusion criteria it was achieved a sample with 116 subjects.

The variables considered in this study were age, height and the spirometric parameters: forced vital capacity (FVC), forced expiratory volume in 1' (FEV1) and FEV1/FVC ratio.

Recently generalised linear models, generalised additive models and generalized linear mixed models have become more widely used than simple linear models. In these the normal distribution for the dependent variable (Y) is replaced by an exponential family of distributions (of which the normal is a special case), and a link function relates de mean value μ, the mean of Y, to the linear predictor. Generalized additive models for location, scale and shape (GAMLSS) extend the above[**?**]. In GAMLSS the exponential family distribution assumption for the response variable (Y) is replaced by a general distribution family that can model both skewness and kurtosis. Thus, GAMLSS offers general linear predictors for all the distribution parameters (μ, σ, ν, τ) and a choice of error distributions. Complex effects of explanatory variables on the dependent variable can be modelled using link functions, which allow the dependent variable to vary smoothly as a function of an explanatory variable. In this work several models were analyzed and the most parsimonious one was selected using the GAIC criterion.
The statistical analysis were conducted using `gamlss` package of R[**?**].

ACKNOWLEDGMENT

This work is partially financed by national funds through FCT Fundação para a Ciência e a

Tecnologia under the project UID/MAT/00006/2013.

References

[1] Miller M.R., Hankinson J., Brusasco V., Burgos F., Casaburi R., Coates A., Crapo R., Enright P., van der Grinten C.P., Gustafsson P., Jensen R., Johnson D.C., MacIntyre N., McKay R., Navajas D., Pedersen O.F., Pellegrino R., Viegi G., Wanger J. (2005). Standardisation of spirometry. *European Respiratory Journal*. 26:319—338.

[2] Miller M.R., Crapo R., Hankinson J., Brusasco V., Burgos F., Casaburi R., Coates A., Enright P., van der Grinten C.P., Gustafsson P., Jensen R., Johnson D.C., MacIntyre N., McKay R., Navajas D., Pedersen O.F., Pellegrino R., Viegi G., Wanger J. (2005). General considerations for lung function testing. *European Respiratory Journal*. 26:153—161.

[3] Vaz Fragoso C.A., McAvay G., Gill T.M., Concato J., Quanjer P.H. and Van Ness P.H. (2014). Ethnic differences in respiratory impairment. *Thorax*, 69:55—62.

[4] Pelligrino R., Viegi G., Bursaco V., Crapo R.O., Burgos F. and Casaburi R. (2005). Interpretive strategies for lung function tests. *European Respiratory Journal*, 26:948—68.

[5] Quanjer P.H., Stanojevic S., Cole T.J., Baur X., Hall G.L., Culver B.H. (2005) ERS global lung function initiative. Multi-ethnic reference values for spirometry for the 3 e 95-yr age range: the global lung function 2012 equations. European Respiratory Journal, 40:1324-–43.

[6] Stasinopoulos, M. Rigby, B. Vlasios, V., Heler, G. and Fernanda, B. (2015). Flexible regression and smoothing the GAMLSS packages in R.

[7] Rigby R.A. and Stasinopoulos D. M (2005). Generalized additive models for location, sacle and shape. *Applied Statistics*, 54:507–554.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

APLICAÇÃO DE PATH ANALYIS NA IDENTIFICAÇÃO DE PREDITORES DA QUALIDADE DE VIDA DE PESSOAS COM DOENÇAS CRÓNICAS

Estela Vilhena[1], José Luís Pais Ribeiro[2] e Denisa Mendonça[3]

[1] EST; 2Ai – Instituto Politécnico do Cávado e do Ave, Barcelos; EPIUnit-ISPUP UP, Porto
[2] Faculdade de Psicologia e Ciências da Educação da UP; William James Center for Research, ISPA– University Institute, Lisbon, Portugal
[3] ICBAS, EPIUnit – ISPUP UP, Porto

RESUMO

Com o objetivo de identificar preditores, simultâneos, a longo prazo da Qualidade de Vida de pessoas com doenças crónicas, foi aplicado o modelo *Path Analysis*. O modelo teórico a testar pressupôs: 1) perceção de estigma, adesão aos tratamentos, otimismo disposicional, afeto positivo e negativo, suporte social são preditores da qualidade de vida; 2) o otimismo exerce um efeito medidor entre a perceção de estigma, adesão aos tratamentos, afeto positivo e negativo, suporte social e a qualidade de vida. Os resultados revelaram um impacto positivo de menor perceção do estigma, do otimismo, do afeto positivo e da adesão aos tratamentos na qualidade de vida destes doentes. Por outro lado, verificou-se ainda que o otimismo exerce um efeito mediador a longo prazo, entre o afeto positivo/negativo e a saúde mental.

Palavras e frases chave: Doença Crónica, *Path Analysis*, Qualidade de Vida.

1. INTRODUÇÃO

O modelo de análise de trajetórias, ou *Path Analysis*, é uma metodologia de análise estatística usada para estudar relações estruturais, permitindo avaliar os efeitos diretos e indiretos entre variáveis [1]. A Qualidade de Vida (QdV) é um constructo composto por um número de fatores que contribuem para o bem-estar de um indivíduo e para o ajustamento a uma determinada doença. Uma doença crónica (DC) é definida como uma doença prolongada, não se resolve espontaneamente e raramente tem cura, sendo responsável por alterações na vida das pessoas [2]. Este trabalho teve como objetivo avaliar um modelo hipotético, que consistiu na análise do impacto simultâneo, a longo prazo, da perceção de estigma, adesão aos tratamentos, otimismo disposicional, afeto positivo e negativo, suporte social nas componentes da QdV (bem-estar geral, saúde física e mental) e concomitantemente, na avaliação do otimismo como efeito mediador [3], num grupo de doentes crónicos.

2. MÉTODOS

O estudo prospetivo envolveu 801 indivíduos com DC (cancro, diabetes, epilepsia, esclerose múltipla, miastenia gravis e obesidade) avaliados em três momentos (T1, T2 e T3) com intervalo

de 8 meses entre os mesmos. Foram considerados para análise os 304 sujeitos que participaram simultaneamente nos três momentos de avaliação. Os critérios de inclusão foram os seguintes: diagnóstico de DC há pelo menos 3 anos; idade superior a 18 anos; nível de escolaridade superior ou igual a 6 anos, vida estabilizada e não apresentar distúrbios psiquiátricos.

Material

Foi aplicado um questionário que incluía um conjunto de variáveis sociodemográficas, perceção de estigma, adesão aos tratamentos, otimismo disposicional, afeto positivo (AP) e negativo (AN), suporte social e as três componentes da QdV.

Perceção de Estigma

Foi usada uma escala desenvolvida por Ribeiro et al. (2009) [4], e ainda em estudo, onde valores mais baixos refletem maior perceção de estigma.

Adesão aos Tratamentos

Desenvolvida por Delgado e Lima [5] foi aplicada a versão Portuguesa, onde valores mais altos significam melhor adesão ao tratamento. A medida mostrou boa consistência interna, 0,74.

Otimismo Disposicional

Avaliado através do Life Orientation Test-Revised. A validação da escala Portuguesa [6] apresenta características semelhantes à versão original, onde pontuações mais elevadas significam um maior grau de otimismo. A versão Portuguesa apresenta um α de Cronbach de 0,71.

Afeto Positivo e Afeto Negativo

O afeto foi avaliado usando o Positive and Negative Affect Schedule, validado para a população Portuguesa [7]. Valores mais elevados do AP indicam mais afeto positivo, ou o grau em que o indivíduo se sente entusiasmado, ativo e alerta. Maior pontuação do NA indica mais afeto negativo, que se reflete em estados aversivos de humor do indivíduo. Verificou-se uma consistência interna de 0,86 para o afeto positivo e 0,89 para a escala do afeto negativo.

Suporte Social

Para a população Portuguesa o apoio social foi avaliado através do Social Support Survey (MOS)106-108 [8]. É composto por quatro subescalas de apoio social distintas, tal como um índice funcional global de apoio social. Todas as subescalas têm mostrado uma forte confiabilidade com um α de Cronbach superior a 0,91.

Qualidade de Vida

Foi utilizada a dimensão bem-estar geral resultante do IQOLA, onde foi encontrado um fator de segunda ordem, com três componentes do SF-36 (bem-estar geral - BEG, saúde física_SF e mental-SM). Cada escala é convertida diretamente numa escala de 0-100 em que 100 representa o nível mais alto. A versão Portuguesa do MOS SF-36 [9] mostra bons níveis de consistência interna (α de Cronbach de 0,70).

Análise Estatística

O modelo *Path Analysis* foi aplicado para testar a qualidade do modelo teórico hipotético. Para testar a adequação do modelo foram usados os índices CFI - *Comparative Fit Index* e o RMSEA

- *Root Mean Error Approximation*. A análise foi efetuada usando o softwae EQS 6.1, com um nível de significância de 0,05.

3. RESULTADOS

Os resultados obtidos revelaram um bom ajustamento do modelo, CFI=0,96 e RMSEA=0,06. As variáveis avaliadas a longo prazo, nomeadamente em T2, exerceram efeitos simultâneos e significativos nas componentes de QdV. Verificou-se que o afeto positivo exerce um impacto positivo e estatisticamente significativo no bem-estar geral e na saúde mental; o afeto negativo e a adesão aos tratamentos exercem um impacto negativo e positivo, respetivamente, estatisticamente significativo nas três componentes da QdV; a perceção de estigma exerce um efeito significativo no bem-estar geral. Por outro lado, os resultados evidenciaram um efeito mediador do otimismo entre o afeto positivo/negativo e a saúde mental.

4. CONCLUSÕES

As técnicas de análise multivariada são hoje em dia cada vez mais aplicadas para dar resposta a questões científicas mais complexas. A *Path Analysis* é uma técnica de análise que teve na sua origem a modelação de relações de causalidade entre variáveis observadas. É usada para testar modelos previamente conjeturados e estudados.

Os resultados encontrados neste estudo sugerem que uma menor perceção do estigma, uma atitude otimista, mais ativa e entusiástica e uma melhor adesão aos tratamentos podem, a longo prazo, facilitar a aceitação do doente à sua nova condição de vida, fator esse, que por sua vez contribuirá para uma menor qualidade de vida.

Referências

[1] Olobatuyi, M. E. (2006). A user's guide to path analysis. UK: University Press of America.

[2] Ribeiro J. (2001) .Qualidade de vida e doença oncológica. In: Durá MRDeE, editor. Territórios da Psicologia Oncológica. Lisboa: Climepsi Editores; 75-98.

[3] Baron, R.M., Kenny, D.A. (1986). The Moderator Mediator Variable Distinction in Social Psychological- Research - Conceptual, Strategic, and Statistical Considerations. Journal of Personality and Social Psychology 51(6), 1173-1182.

[4] Pais-Ribeiro J, Silva I, Abreu M, Costa N, Cardoso H, Venâncio C. (2009) Stigma and quality of life of obese women – preliminary Study. The European Journal of Obesity, 2 (sup 2): 244

[5] Delgado A, Lima M. (2001) Contributo para a avalição concorrente de uma medida de adesão aos tratamentos. Psicologia, Saúde & Doenças, 2 (2): 81-100.

[6] Pais Ribeiro J, Pedro L. (2006) Contribuição para a análise psicométrica e estrutural da escala revista de avaliação do optimismo (escala de orientação para a vida revista-EOR-R) em doentes com esclerose múltipla. In: Leal I, Pais Ribeiro J, Neves S, editors. Actas do 6º Congresso Nacional de Psicologia da Saúde;. p. 133-9.

[7] Galinha IC, Pais Ribeiro JL. (2005) Contribuição para o estudo da versão portuguesa da Positive and Negative Affect Schedule (PANAS): II – Estudo psicométrico. Análise Psicológica, 2(XXIII): 219-29.

[8] Pais-Ribeiro J, Ponte AC. (2009) Propriedades métricas da versão portuguesa da escala de suporte social do MOS (MOS Social Support Survey) com idosos. Psicologia, Saúde & Doenças, 10(2): 163-74

[9] Ferreira, P. (2000). Criação da versão portuguesa do MOS SF-36: Parte II - Testes de validade. Acta Médica Portuguesa, 13, 55-66.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

DETERMINAÇÃO DA COMPOSIÇÃO CORPORAL EM JOVENS ADULTOS - AVALIAÇÃO DA REPRODUTIBILIDADE ENTRE PROTOCOLOS ECOGRÁFICOS E IDENTIFICAÇÃO DE PREDITORES DE MASSA GORDA TOTAL

Mário Monteiro[1,] João Paulo de Figueiredo[2], Sandra Assunção[1], Rute Santos[1,], António Figueiredo[3],

[1] Departamento de Imagem Médica e Radioterapia, Escola Superior de Saúde de Coimbra (ESTeSC), Instituto Politécnico de Coimbra

[2] –Departamento de Ciências Complementares (Estatística e Epidemiologia), Escola Superior de Saúde de Coimbra (ESTeSC), Instituto Politécnico de Coimbra

[3] – Faculdade de Ciências do Desporto e Educação Física, Universidade de Coimbra

RESUMO

Introdução: A avaliação da composição corporal permite um estudo da condição física e dos diferentes componentes corporais. Esta possibilita a determinação das percentagens de massa magra e massa gorda. Existem diferentes métodos de avaliação da composição corporal. A DEXA (dual – energy X-Ray absorptiometry) e a Ecografia além de serem empregues regularmente no diagnóstico clínico são métodos indiretos de avaliação da composição corporal. **Objetivos**: comparação de dois diferentes protocolos ecográficos para a avaliação do tecido celular subcutâneo e relacionar os dados obtidos através da DEXA bem como a determinação de preditores de massa gorda. **Material e Métodos**: Vinte indivíduos (10 do género masculino e 10 do género feminino) com idades compreendidas entre os 19 e os 24 anos foram submetidos a um conjunto de avaliações antropométricas seguindo-se duas avaliações ecográficas (protocolos distintos) e por fim foram submetidos a DEXA. **Resultados**: A reprodutibilidade dos protocolos ecográficos utilizados apresentou valores elevados de CCI na ordem dos 0,9. O protocolo 1 mostrou ser um bom preditor na avaliação da massa gorda, apresentando uma correspondência de 82% ($p< 0,0001$) com a DEXA, já protocolo 2 apresentou uma correspondência de 85%. **Conclusões**: Com este estudo foi possível desenvolver duas equações de regressão linear para a avaliação da massa gorda, com base nos dois protocolos em estudo e na DEXA.

Palavras e frases chave: Ecografia, DEXA, Reprodutibilidade, Massa Gorda Total.

1. INTRODUÇÃO

A avaliação da composição corporal (CC) permite um estudo da condição física e dos diferentes componentes corporais. Possibilita a determinação das percentagens de massa magra e massa gorda. A massa gorda compreende todos os lípidos do tecido adiposo e dos outros tecidos. Enquanto, a massa magra (massa livre de gordura) inclui os resíduos químicos e todos os outros tecidos isentos de gordura, tais como a água, os ossos, os músculos, o tecido conjuntivo e os órgãos internos [1,2]. Vários autores confirmam que o mais importante é conhecer a forma como a gordura se distribui no corpo[2,3]. Existem diferentes métodos de avaliação da CCe estes podem ser agrupados em três níveis diferentes de análise: os diretos, os indiretos e os duplamente indiretos. No caso da dual – *energy X-Ray absorptiometry* ou densitometria radiológica de dupla energia (DEXA) e da Ecografia, além de serem empregues regularmente no diagnóstico clínico são métodos indiretos de avaliação da composição corporal [2,3]. A Ecografia, ou Ultrassonografia, é uma técnica de não invasiva, que não utiliza radiação ionizante e não tem quaisquer restrições para a sua realização. Geralmente a ecografia fornece imagens claras das estruturas e que podem ser facilmente identificáveis permitindo avaliar a espessura do músculo e da gordura em diferentes regiões do corpo [1,4,5,6]. A DEXA e a Ecografia são dois métodos de imagem que nos permitem fazer a quantificação da massa

gorda e da massa magra. Na Ecografia essa avaliação é feita através da medição da espessura do tecido subcutâneo e/ou do músculo, recorrendo-se posteriormente a cálculos para a obtenção dos valores pretendidos [1,2]. A DEXA possibilita uma avaliação corporal total ou por segmentos e dá informação sobre os diferentes componentes da CCsendo considerada um método *gold standard* devido à sua boa precisão e reprodutibilidade [7,8,9,10]. Propusemos como objetivo principal deste estudo a comparação de dois diferentes protocolos ecográficos para a avaliação do tecido celular subcutâneo e através desta avaliação calcular a quantidade de massa gorda total, tendo como referência a DEXA.

2. MATERIAL E MÉTODOS

Integraram no estudo 10 jovens adultos do género masculino e 10 do género feminino com idades compreendidas entre os 19 e os 24 anos com Índice de Massa Corporal dentro dos parâmetros normais (18,5-24,9Kg/m^2) e que foram submetidos a um conjunto de avaliações, nomeadamente duas avaliações ecográficas e uma avaliação por DEXA. Para avaliação antropométrica utilizou-se a balança digital *Marsden modelo M-120 GP Column Scale* (capacidade 250Kg), com estadiómetro acoplado (0 - 200cm). Posteriormente, foi realizada uma DEXA (*Densitometria Bifotónica* [DXA] *Lunar* iDXA da *General Electrics Healthcare - software GE enCORE Versão* 13.60.300) para avaliação da CC de corpo inteiro. De seguida, para a avaliação ecográfica utilizou-se um ecógrafo portátil *(LOGIQe, General Electrics Healthcare*), com uma sonda linear (7-12 MHz), tendo sido os parâmetros ecográficos mantidos durante todas as avaliações. As imagens adquiridas, e posteriormente guardadas, foram analisadas através do *software Image J* (*National Institutes of Health, Bethesda,* MD, EUA). Todas as avaliações e análise dos dados foram realizadas pelo mesmo examinador. Nas imagens ecográficas a espessura do tecido adiposo foi medida entre o limite mais profundo da pele e a fáscia superficial muscular. Ambas as avaliações e medições de Eco e de DEXA foram realizadas no mesmo dia. Os protocolos em avaliação foram o protocolo usado pelo ISAK (*International Standards for Anthropometric Assessment*) para a medição de pregas cutâneas (protocolo 1) e o protocolo sugerido pelo autor Muller W. et al (protocolo 2) [11,19].

Ao nível dos modelos estatísticos aplicaram-se os seguintes testes: t-Student para amostras emparelhadas, T de Wilcoxon, Coeficiente de Correlação Linear de Pearson, Coeficiente de Variação, Coeficiente de Correlação Intra-Classe (CCI) e o Método de Análise de Regressão Linear Simples/Múltipla.

3. RESULTADOS

Relativamente à reprodutibilidade dos protocolos ecográficos utilizados, verificaram-se valores elevados de CCI na ordem dos 0,90.

Perante os dois protocolos, inicialmente foram estimados os valores dos parâmetros compatíveis dos dois protocolos, sendo eles nomeadamente: o músculo do tricípite, abdómen, coxa e o músculo gastrocnémio. Segundo os resultados apresentado na tabela 1, observaram-se diferenças significativas, entre os dois protocolos, para a avaliação do *Tricípite*, a nível da avaliação do *Abdómen Inferior* e bem como a nível da estimativa *média dos valores do Abdómen Inferior e superior* (tabela 1).

N=20	Média	Desvio Padrão	*p*
Protocolo 1 - tricípite	0,80	0,44	<0,0001
Protocolo 2 - tricípite	0,57	0,36	
Protocolo 1 - Abdómen	1,04	0,56	0,37
Protocolo 2 – Abdómen superior	1,02	0,61	
Protocolo 1 - Abdómen	1,04	0,56	<0,0001
Protocolo 2 – Abdómen inferior	1,55	0,71	
Protocolo 1 - Abdómen	1,04	0,56	<0,0001
Protocolo 2 – Abdómen Superior/inferior (*)	1,29	0,64	
Protocolo 1 - Coxa	0,88	0,41	0,23
Protocolo 2 – Coxa anterior	0,91	0,45	
Protocolo 1 - Gastrocnémio	0,55	0,26	0,96
Protocolo 2 – Gastrocnémio	0,55	0,26	

Tabela 1: Comparação dos valores estimados dos parâmetros compatíveis dos dois protocolos. Teste: t-Student e T de Wilcoxon; (*) – Estimativa Média entre Abdómen Superior e Inferior

No caso do *Abdómen Inferior* e da estimativa *Média entre Abdómen Superior e Inferior* apresentou valores médios superiores nestas regiões comparativamente com o protocolo 1 para o *Abdómen*. Contrariamente, no parâmetro de avaliação do *Tricípite* o protocolo 2 apresentou valores mais baixos que o protocolo 1. Nos restantes locais de medição não se verificam diferenças significativas ($p > 0,05$).

Seguidamente, foi avaliada a correlação entre estas regiões corporais dos dois protocolos, com recurso a análise de correlação linear, coeficientes de variação e bandas de confiança a 95% para a média. No que diz respeito à avaliação do *Tricípite* ambos os protocolos apresentaram uma elevada correlação (concordância positiva) de 86% (ρ = 0,930). Recorrendo à mesma estratégia de análise para os outros parâmetros de avaliação verificou-se que no caso do *Abdómen* e do *Abdómen Superior* o padrão de correlação mostrou ser ligeiramente inferior (Coeficiente de variação: 78% para ρ = 0,880) comparativamente à avaliação do Tricípite, bem como no *Abdómen* e do *Abdómen Inferior* (Coeficiente de variação: 77% para ρ = 0,880) e no *Abdómen* e estimativa *média dos valores do Abdómen Inferior e superior* (Coeficiente de variação: 81% para ρ = 0,90). No caso dos dois últimos parâmetros de avaliação verificou-se que a *Coxa* e a *Coxa Anterior* apresentaram uma concordância de 91% (ρ = 0,95) e os *Músculos Gastrocnémio* dos dois protocolos apresentaram 88% de concordância (ρ = 0,94).

Assim, pode-se concluir que estes parâmetros comuns a ambos os protocolos mostraram uma boa correlação entre si. Propusemos avaliar, de seguida, os preditores que melhor poderiam explicar a presença de massa gorda total quando as mesmas foram avaliadas por DEXA.

Protocolos	Variáveis preditores	$\widehat{\beta}$ ($\hat{\sigma}$)	β	*p*	R^2ajustado	$R^2\Delta$	*p*
1	Supra-espinhal	18,47 (3,73)	0,83	<0,0001	0,77	0,82	<0,001
	Gastrocnémio	6,57 (7,14)	0,22	0,37			
	Tricípite	4,34 (3,64)	0,25	0,25			
	Bicípite	-25,07 (13,87)	-0,37	0,09			
2	Oblíquo externo	16,69 (2,51)	0,83	<0,0001	0,81	0,85	<0,001
	Braquio-radial	26,14 (12,93)	0,35	0,06			
	Coxa anterior	-8,13 (4,87)	-0,47	0,12			
	Coxa lateral	3,75 (1,55)	0,55	0,030			

Tabela 2 Regressão linear múltipla entre a DEXA e o protocolo 1 e 2. Variável dependente = valores de massa gorda fornecidos pela DEXA (Kg) $\widehat{\beta}$ – Coeficiente de regressão não estandardizado; $\hat{\sigma}$ - Erro padrão do coeficiente de regressão não estandardizado; *p* – valor de significância.

Após a análise dos valores estatísticos do protocolo 1, tendo em conta o conjunto de preditores, concluiu-se que este é um bom preditor na avaliação da massa gorda, apresentando uma correspondência de 82% ($R^2\Delta = 0,82$; $p < 0,0001$) com a DEXA. Perante a avaliação dos coeficientes de regressão parciais do protocolo 1, pode constatar-se que o parâmetro de avaliação Supra-espinhal explicou positivamente a variação dos valores do tecido celular subcutâneo total em 83% (p<0,0001). A Massa Gorda Total (MGT) pode ser predita segundo a equação apresentada a seguir:

$$MGT_{p1} = 6,99 + (18,47 \times supra - espinhal) + (6,57 \times \text{Gastrocnémio}) + (3,34 \times \text{Tricípite}) + (-25,07 \times \text{Bicípite})$$

Tal como ocorre com o protocolo 1, na análise dos valores obtidos do protocolo 2, tendo em conta as variáveis preditores, constatou-se que este apresentou um efeito explicativo de 85% na avaliação da massa gorda total. Perante os valores obtidos na avaliação dos coeficientes de regressão parciais pode-se constatar que Oblíquo Externo (83%; $p < 0,0001$) e Coxa lateral (55%; $p = 0,03$) predisseram positivamente a variação do tecido celular subcutâneo total. Estes valores permitiram gerar a seguinte fórmula:

$$MGT_{p2} = 3,45 + (16,69 \times Oblíquo\ externo) + (26,14 \times \text{Bráquio radial}) + (-8,13 \times \text{Coxa Anterior}) + (3,75 \times \text{Coxa Lateral})$$

Esta prevê a massa gorda total (MGT), segundo a constante e as variáveis preditores do protocolo 2 anteriormente apresentadas.

4. DISCUSSÃO e CONCLUSÕES

Perante os resultados obtidos do coeficiente de correlação do protocolo 1 e 2 com a DEXA conclui-se que os seus valores encontram-se fortemente correlacionados positivamente (ecografia e DEXA) o que vai de acordo com os resultados obtidos num estudo realizado por Pineu J. et al [13]. Sendo a ecografia um bom preditor para a determinação do tecido celular subcutâneo, bem como para a determinação da massa gorda, como é referido num outro estudo de Pineu J. et al, este é um método de avaliação que tem vindo a ser mais utilizado para o estudo da gordura subcutânea [14]. Como é referido por muitos autores na literatura, a avaliação ecográfica não requer um elevado grau de experiência, é fácil de ser realizada, é segura, os seus resultados são imediatos e envolve um equipamento que poderá ser portátil [6,13,14, 15]. No entanto, como é referido por Duz S. et al existem poucos estudos na literatura que tenham desenvolvido equações de

regressão para determinar a massa gorda total, partindo de medições ecográficas e correlacionando com os valores fornecidos pela DEXA [16]. Partindo deste pressuposto, Muller W. et al defende a ecografia como um bom método de avaliação da gordura subcutânea, contudo afirma que o protocolo do ISAK, usado para a avaliação de pregas cutâneas, que é maioritariamente usado, presenta baixa reprodutibilidade. Refere ainda que é necessário o uso de um protocolo que apresente apenas regiões de interesse para a avaliação da massa gorda [17,18]. Face ao exposto foi possível desenvolver duas equações de regressão para a avaliação do tecido gordo total com base nos dois protocolos em estudo, ambas são preditores para a avaliação da massa gorda uma vez que os protocolos correspondentes apresentaram uma correspondência com a DEXA superior a 80%. Perante os resultados apresentados é possível concluir que o protocolo sugerido pelo autor Muller W. et al apresenta um resultado ligeiramente superior ao protocolo do ISAK devendo, por isso, ser o protocolo a adotar.

Referências

[1]. Leite, M. (2004). Métodos de avaliação da composição corporal. Faculdade de Ciências da Nutrição e Alimentação da Universidade do Porto. Retrieved from URL: https://repositorio-aberto.up.pt/bitstream/10216/54643/5/103136_04-57T_TL_01_P.pdf

[2]. Vieira, F., Fragoso I. (2006). Composição Corporal. In: Morfologia E Crescimento. 2nd ed. Lisboa.

[3]. Murmann, B. (2006). A avaliação da composição corporal - a medição de pregas adiposas como técnica para a avaliação da composição corporal. Ieee Micro.

[4]. Sant SL., Elo S., Franceschini C.C. (2009). Métodos de avaliação da composição corporal em crianças;27(3):315-321.

[5]. Monteiro, A.B. (1984). Artigo de Revisão análise da composição corporal : uma revisão de métodos analysis of the body composition. A review of methods.

[6]. Utter A.C., Hager M.E. (2008). Evaluation of ultrasound in assessing body composition of high school wrestlers. Med Sci Sports Exerc.40(5):943-949. doi:10.1249/MSS.0b013e318163f29e.

[7]. Company J., Ball S. (2010). Body Composition Comparison : Bioelectric Impedance Analysis with Dual-Energy X-Ray Absorptiometry in Adult Athletes:186-201. doi:10.1080/1091367X.2010.497449.

[8]. Mello M.T. De, Dâmaso A.R., Antunes H.K.M., et al. (2005). Avaliação da composição corporal em adolescentes obesos : o uso de dois diferentes métodos.11:267-270.

[9]. Bilsborough J.C., Greenway K., Opar D., Livingstone S., Cordy J., Coutts A.J. (2014). The accuracy and precision of DXA for assessing body composition in team sport athletes:37-41. doi:10.1080/02640414.2014.926380.

[10]. Sciences S., Gobbo L., Gon E. (2013). Body composition in taller individuals using DXA : a validation study for athletic and non- athletic populations.

[11]. Müller W., Lohman T.G., Stewart A.D., et al. (2016). Subcutaneous fat patterning in athletes : selection of appropriate sites and standardisation of a novel ultrasound measurement technique : ad hoc working group on body composition , health and performance , under the auspices of the IOC Medical Commission:45-54. doi:10.1136/bjsports-2015-095641.

[12]. International Society for the Advencement of Kinanthropometry (2001). International Standards for Anthropometric Assessment. Retrieved from URL: http://www.ceap.br/material/MAT17032011184632.pdf

[13]. Pineau J., Bocquet M., Peres G. (2010) Ultrasound Measurement of Total Body Fat in Obese Adolescents. Ann Ntrition Metab;56.

[14]. Pineau J., Filliard J.R., Bocquet M. (2009). Ultrasound Techniques Applied to Body Fat Measurement in Male and Female Athletes.44(2):142-147.

[15]. Wagner D.R. (2013). Ultrasound as a Tool to Assess Body Fat.

[16]. Duz S., Kocak M., Korkusuz F. (2009) Evaluation of body composition using three different methods compared to dual-energy X-ray absorptiometry. Eur J Sport Sci. 9:181-190.

[17]. Müller W., Horn M., Fürhapter-rieger A., et al. (2013). Body composition in sport : interobserver reliability of a novel ultrasound measure of subcutaneous fat tissue:1036-1043. doi:10.1136/bjsports-2013-092233.

[18]. Müller W., Horn M., Fürhapter-rieger A., et al. Body composition in sport : a comparison of a novel ultrasound imaging technique to measure subcutaneous fat tissue compared with skinfold measurement. 2013:1028-1035. doi:10.1136/bjsports-2013-092232.

[19]. Müller W., Maughan R.J. (2013). The need for a novel approach to measure body composition : is ultrasound an answer ? 2013;47(16):1001-1002. doi:10.1136/bjsports-2013-092882.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

PERMUTATION DISTRIBUTIONS FOR PATTERN CLASSIFICATION IN NEUROIMAGING

Mohammed S. Al-Rawi[1], Adelaide Freitas[2], João V. Duarte[3] and Miguel Castelo Branco[3,4]

[1]Institute of Electronics and Telematics Engineering of Aveiro (IEETA), University of Aveiro, Portugal
[2]Department of Mathematics & CIDMA, University of Aveiro, Portugal
[3]Visual Neuroscience Laboratory & IBILI, Faculty of Medicine, University of Coimbra, Portugal
[4]Institute of Nuclear Sciences Applied to Health, University of Coimbra, Portugal

ABSTRACT

Permutation tests have been used to estimate the statistical significance of classifiers in neuroimaging where the number of observations is relatively small and their dimensionality is high. In a two-class comparison, a binary classifier is trained in order to capture differences between the two groups and thus to label new examples better than by chance. The misclassification error of the classifier is the statistic used to measure how dissimilar the two classes are. The misclassification error represents a sum of Bernoulli variables and the empirical distribution of the average misclassification error is typically deployed by repetitions of cross-validation procedures, that are centered around chance-level, for a huge of quantity of permutations of the labels in the original data set. Hence, we can suspect the average of the misclassification error could be approximately modeled by a gaussian distribution, under the assumption of independence. However, there is correlation among the cross-validation folds. In this work, we discuss a simulation study carried by us (Al-Rawi et al., 2017) using functional magnetic resonance imaging data to evaluate the deviation of the permutation distribution of average misclassification errors from gaussian distribution using Anderson-Darling test.

Keywords and key sentences: Permutation test, cross-validation, Anderson-Darling test.

1. INTRODUCTION

Classification techniques on brain images acquired via functional Magnetic Resonance Imaging (fMRI) have been used for decoding mental states from patterns of brain activities in humans. Permutation tests usually use the test error as a data set statistic to estimate the p-value(s) by measuring the dissimilarity between two or more populations.
An exhaustive simulation study using several binary classifiers was carried out by us (Al-Rawi et al., 2017) using a real fMRI data set collected while human subjects responded to visual stimulation paradigms. Two scrambling schemes are evaluated: the first based on

permuting both the training and the testing sets, and the second based on permuting only the testing, such that the classifier model is obtained by the non-permuted set. In this work, we emphasize the main contributions of that paper, observing that a normal distribution can not adequately fit to permutation distributions (PDs) most of the times. However, normal distribution tends to be quite well acceptable when mean classification accuracies averaged over a set of independent classifiers is considered.

2. PERMUTATION DISTRIBUTION

When running permutation testing procedures to estimate the classification significance, a classifier is used to classify the data of each permutation to build an empirical PD. A PD represents an empirical estimate of the cumulative distribution of the error test or the accuracy of the classifier under the null hypothesis of independence between the data and the labels. The error test can be estimated using cross-validation in each iteration of the permutation procedure. The theoretical null distribution of the average misclassification error (or percentage of incorrectly estimated labels) in each cross-validation fold is proportional to a binomial distribution, under the independence of the the cross-validation fold, which is not true. However, several practice situations have widely accepted that possible dependence features associated with non-overlapping folds or cross-validation could be negligible. Hence, it is commonly accepted that every PD is normally distributed and centered on theoretical chance-level. But, it can not be true!
Thus, investigating the deviation of PDs to normal law is a vital topic and could facilitate obtaining enhanced mental state decoding techniques. The motives behind measuring the normality of PDs are: comparing different classification models, inferring the dependence across fMRI runs, finding enhanced data partitioning schemes, spotlighting the method of cross-validation and whether it can be considered as inadequate partitioning procedure. To investigate the shapes of PDs, Al-Rawi el at (2017) built several PDs via classifying fMRI data taken when the subjects responded to visual stimulation paradigms. Anderson-Darling test (AD-test) was used to infer whether a PD is normal or not, i.e., to test the null hypothesis that the random sample of the test error values of a classifier, or a system of classifiers, is generated by a normal distribution.

3. CONCLUSIONS

Among the six classifiers considered by Al-Rawi et al. (2017), the simulation study showed that the permutation distribution of the percentage of misclassification based on single single classifier does not tend to have normal distribution. Nevertheless, for each of the scrambling schemes (scramble the training and the testing set, and scramble only the testing set), a performance metric based on the mean percentage of misclassification of a group of five (or more) classifiers can result in permutation distributions which can be assumed to be gaussian. Unpredictably, scrambling both the training and the testing set did not provoke PDs to approximate to normal law. It seems that scrambling can have different inter and intra stimulus dependency in the permutation procedure.

ACKNOWLEDGMENT

This work was supported by Portuguese funds through the CIDMA - Center for Research & Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), within project designed by UID/MAT/04106/2013, and by Quadro de Referência Estratégico Nacional (national strategic reference framework, QREN) under the Mais Centro initiative: CENTRO-07-ST24-FEDER-00205, PEst/C/SAU/3282/2013, COMPETE FCOMP-01-0124-FEDER-022690.

References

[1] Al-Rawi, M.S., Freitas, A., Duarte, JV., Cunha, J.P., Castelo-Branco, M. (2017) Permutations of functional magnetic resonance imaging classification may not be normally distributed. *Statistical Methods in Medical Research* Vol. 26, 6, 2567–2585.

[2] Etzel J.A., Gazzola V., Keysers C.(2009) An introduction to anatomical ROI-based fMRI classification analysis. *Brain Research* 1282, 114–125.

[3] Golland, P., Fischl, B. (2003) Permutation tests for classification: Towards statistical significance in image-based studies In Taylor, C., Noble, J.A. (eds): *Information Processing in Medical Imaging, Proceedings* 330-341, Springer-Verlag Berlin: Berlin.

III Encuentro Luso-Galaico de Biometría
Aveiro, 28 - 30 de Junio de 2018

UN MODELO DE OPTIMIZACIÓN CONTINUA MULTIOBJETIVO PARA PLANIFICACIÓN FORESTAL

Jose M. González-González[1], Miguel E. Vázquez-Méndez[2] y Ulises Diéguez-Aranda[1]

[1]Unidade de Xestión Forestal Sostible, Departamento de Enxeñaría Agroforestal, Universidade de Santiago de Compostela. EPS de Enxeñaría, Rúa Benigno Ledo, Campus Terra, 27002 Lugo, Spain.
[2]Departamento de Matemática Aplicada, Universidade de Santiago de Compostela. EPS de Enxeñaría, Rúa Benigno Ledo, Campus Terra, 27002 Lugo, Spain.

RESUMEN

En este trabajo se propone una formulación novedosa para planificar la gestión forestal de un monte. El modelo consiste en un problema de optimización multiobjetivo formulado con variables continuas, que puede abordarse con técnicas de optimización diferenciable. La bondad del modelo se ilustra utilizándolo para la toma de decisiones en un monte de *Eucalyptus globulus* Labill. ubicado en el norte de Galicia.

Palabras y frases clave: Optimización diferenciable. Ordenación de montes. Constancia de rentas. Frente Pareto.

1. INTRODUCCIÓN

Rentabilidad y sostenibilidad son dos aspectos fundamentales de la gestión forestal de un monte. La rentabilidad está directamente relacionada con los beneficios (B) que proporciona esa gestión, mientras que la sostenibilidad recoge diferentes aspectos entre los que destaca, por ejemplo, la constancia de rentas (C). Consecuentemente, el marco natural para planificar la gestión forestal de un monte (ordenación) es el de la optimización multiobjetivo. Centrándonos únicamente en beneficios y constancia de rentas, la optimización en la ordenación de montes consiste en estudiar y resolver el problema

$$\text{maximizar } \mathbf{J} = (B, C). \tag{1}$$

Aunque la idea es muy simple, el hecho de que los objetivos considerados entren en conflicto hace necesario incluir la toma de decisiones en la resolución del problema (1), y convierten el concepto de *ordenación forestal óptima* en algo subjetivo. Una primera alternativa es utilizar una estrategia *a priori* para abordar el problema, sustituyendo (1) por un problema con un único objetivo. Así, siguiendo a [4], muchos autores han considerado como ordenación forestal óptima aquella que maximiza beneficios, garantizando una diferencia de volúmenes cortados entre periodos (modo habitual de medir la constancia de rentas) menor que un cierto

valor dado. Otros autores (ver, por ejemplo, [3]) proponen obtener todo el frente Pareto del problema (1) para, posteriormente, abordar la toma de decisiones. En cualquier caso, cuando se exige que a cada rodal se le aplique un único programa selvícola, el problema (1) se suele formular como un problema combinatorio, tratando las variables de decisión como variables discretas, a pesar de que éstas son, por su propia naturaleza, claramente continuas.
En este trabajo se presenta una manera alternativa de medir la constancia de rentas, que consiste en ver cuánto se ajusta el volumen de madera cortado a una función constante en el tiempo. Se muestra cómo de este modo se consigue que la función C admita derivadas parciales con respecto a las variables de diseño, lo que nos lleva a tratar el problema (1) con variables continuas y abordarlo con técnicas de optimización diferenciable. Para ilustrar este hecho se toma como estudio de caso un monte de *Eucalyptus globulus* Labill. en Xove (norte de Galicia), y se combina el método de ϵ-restricciones con un algoritmo de programación cuadrática sucesiva (SQP) para obtener el frente Pareto del problema.

2. ORDENACIÓN FORESTAL ÓPTIMA: UNA FORMULACIÓN DERIVABLE

Se entiende por rodal un espacio forestal con características de masa y estación (clima, suelo y fisiografía) homogéneas. Se considera entonces un monte formado por la agregación de N_R rodales, y se busca su ordenación para un horizonte de planificación de T años. El objetivo es determinar, para cada rodal (rodal j), el número de claras a realizar (N_C^j), la relación de extracción ($R_i^j \in (0,1]$), intensidad ($I_i^j \in (0,1)$) e instante ($t_i^j \in [0,T]$) en el que debe aplicarse cada una de ellas (clara i), así como el instante de la corta final ($t^j \in [0,T]$). La variable de decisión a nivel rodal resulta

$$\mathbf{u}^j = (t_1^j, I_1^j, R_1^j, \ldots, t_{N_C^j}^j, I_{N_C^j}^j, R_{N_C^j}^j, t^j) \in \mathbb{R}^{3N_C^j+1},$$

y, consecuentemente, la variable de decisión a nivel monte tiene dimensión $d = N_R + 3\sum_{j=1}^{N_R} N_C^j$, y viene dada por

$$\mathbf{u} = (\mathbf{u}^1, \ldots, \mathbf{u}^{N_R}) \in \mathbb{R}^d.$$

Los beneficios de un monte se obtienen como la suma de los beneficios que proporciona cada uno de los rodales. Los beneficios de un rodal pueden medirse en términos del volumen de madera extraído, del valor actual neto (VAN), del valor esperado del suelo (VES), etc. Dependiendo de la especie y de la zona geográfica, en las últimas décadas se han ido obteniendo expresiones explícitas que proporcionan los valores de esos indicadores en función de la variable de decisión del rodal (ver, por ejemplo, [1]). Se parte pues de que, para cada rodal j, se conoce una función $B^j(\mathbf{u}^j)$ que da los beneficios de ese rodal, de modo que los beneficios del monte vienen dados por

$$B(\mathbf{u}) = \sum_{j=1}^{N_R} B^j(\mathbf{u}^j). \tag{2}$$

La constancia de rentas suele tratarse en términos de volúmenes cortados, y para ello se pueden suponer conocidas las funciones $V_i^j(\mathbf{u}^j)$ y $V^j(\mathbf{u}^j)$, que proporcionan los volúmenes de madera del rodal j que se obtienen en la clara i y en la corta final, respectivamente. Para medir la constancia de rentas resulta habitual dividir el horizonte de planificación en periodos de idéntica duración, y buscar que los volúmenes cortados en los diferentes periodos sean similares. Esa división es totalmente artificial y resulta más útil establecer una cierta constancia (regularidad) del volumen cortado durante todo el horizonte de planificación. Teniendo en cuenta que un rodal no se corta en un único instante, sino que la actuación dura un pequeño intervalo de tiempo (días o incluso semanas), teóricamente la regularidad total podría alcanzarse si el monte fuese suficientemente grande. Regularizar rentas es tratar de que

el volumen cortado se ajuste a una función constante en todo el horizonte de planificación, y como indicador de esa constancia puede elegirse cualquier métrica que mida la bondad del ajuste. Con esa idea en mente, consideramos la función $V_{ac}(\mathbf{u},t)$ que proporciona el volumen cortado acumulado en el instante $t \in [0,T]$, si se sigue la planificación definida por la variable $\mathbf{u}$. Esa función viene dada por

$$V_{ac}(\mathbf{u},t) = \sum_{j=1}^{N_R} V^j(\mathbf{u}^j) H_{t^j}(t) + \sum_{j=1}^{N_R} \sum_{i=1}^{N_C^j} V_i^j(\mathbf{u}^j) H_{t_i^j}(t),$$

donde $H_{\bar{t}}(t)$ es la función salto unidad en $\bar{t}$ definida por: $H_{\bar{t}}(t) = \begin{cases} 0 & \text{si } t < \bar{t}, \\ 1 & \text{si } t \geq \bar{t}. \end{cases}$

Para medir la calidad del ajuste entre esta función y el volumen acumulado teórico (el que se correspondería con un volumen cortado constante), en este trabajo proponemos utilizar la métrica asociada a la norma $L^2(0,T)$. Esto nos lleva a definir la constancia de rentas (a maximizar) como

$$C(\mathbf{u}) = -\int_0^T \left(V_{ac}(\mathbf{u},t) - b(\mathbf{u})t\right)^2 \, dt, \tag{3}$$

donde $b(\mathbf{u}) = \dfrac{V_{ac}(\mathbf{u},T)}{T} = \dfrac{\sum_{j=1}^{N_R} \left(V^j(\mathbf{u}^j) + \sum_{i=1}^{N_C^j} V_i^j(\mathbf{u}^j) \right)}{T}$.

Recientemente [1] se ha observado que las funciones $B^j(\mathbf{u}^j)$, $V_i^j(\mathbf{u}^j)$ y $V^j(\mathbf{u}^j)$ son diferenciables (clase C^∞) y, consecuentemente, la función $B(\mathbf{u})$ también lo es. La función $C(\mathbf{u})$ definida en (3) tiene ciertas propiedades de regularidad y, por ejemplo, si no se realizan claras ($N_C^j = 0$, $\mathbf{u} = (t^1, \ldots, t^{N_R})$ y la notación es mucho menos engorrosa), se tiene que:

Teorema 1 *La función C dada por (3) es continua y admite derivadas parciales continuas respecto a la variable t^j en todos los puntos $\bar{\mathbf{u}} = (\bar{t}^1, \ldots, \bar{t}^{N_R})$ que cumplan $\bar{t}^j \neq \bar{t}^k$ para todo $k \neq j$, verificándose además que:*

$$\begin{aligned} \frac{\partial C}{\partial t^j}(\bar{\mathbf{u}}) = & \quad 2\, V^j(\bar{t^j}) \left(V_{ac}(\bar{\mathbf{u}}, \bar{t}^j) - \frac{V^j(\bar{t}^j)}{2} - b(\bar{\mathbf{u}})\bar{t}^j \right) \\ & - \quad 2\, V^{j\prime}(\bar{t}^j) \int_0^T \left(V_{ac}(\bar{\mathbf{u}}, t) - b(\bar{\mathbf{u}})t \right) \left(H_{\bar{t}^j}(t) - \frac{t}{T} \right) dt. \end{aligned}$$

Este resultado permite el uso de métodos tipo gradiente en la resolución del problema (1). Aun así, es preciso tener en cuenta que la función C dada por (3) puede presentar máximos locales, de manera que esos métodos deben combinarse con estrategias de optimización global.

3. ESTUDIO DE CASO: UN MONTE DE *Eucalyptus globulus* EN EL NORTE DE GALICIA

Consideramos un monte de *E. globulus* ubicado en el Concello de Xove, al norte de Galicia, formado por $N_R = 51$ rodales. Para cada rodal disponemos de la fecha de plantación, la superficie (A^j, en hectáreas) y mediciones, en un determinado instante, de la altura dominante, el número de árboles y el área basimétrica. Con esos datos utilizamos un modelo dinámico de crecimiento apropiado para rodales regulares de *E. globulus* en Galicia (ver [2]) que nos proporciona, en cada instante t (años), la altura dominante ($H^j(t)$, en metros), el número de árboles por hectárea ($N^j(t)$) y el área basimétrica ($G^j(t)$, en m^2/ha). Admitimos que no se realizan claras (N_C^j=0) y, de este modo, el volumen de madera ($V^j(t)$, en m^3) existente en cada rodal a lo largo del tiempo viene dado por (ver [2])

$$V^j(t) = A^j \left(0.6234\, H^j(t)^{0.8642} N^j(t)^{\text{-}0.05978} G^j(t)^{1.108} \right).$$

El beneficio (VAN) del monte se obtiene como

$$B(\mathbf{u}) = p \sum_{j=1}^{N_R} \frac{V^j(t^j)}{(1{+}0.01r)^{t^j}} - c_r \sum_{j=1}^{N_R} \frac{A^j}{(1{+}0.01r)^{t^j}},$$

donde r (%) es la tasa de interés anual, p (€/m^3) es el precio esperado de venta de la madera y c_r (€/ha) es el coste de regeneración. En esta situación, para facilitar la planificación del monte en un horizonte de $T = 20.5$ años, buscamos el frente Pareto del problema (1) con un método de ϵ-restricciones, utilizando un algoritmo SQP con multiarranque para resolver cada uno de los problemas de optimización necesarios. En la Figura 1 se muestran (a) la forma del frente y las soluciones (instantes de corta y volúmenes acumulados correspondientes) para tres de los puntos del mismo: (b) el mejor desde el punto de vista económico, (c) uno intermedio y (d) el mejor en cuanto a constancia de rentas.

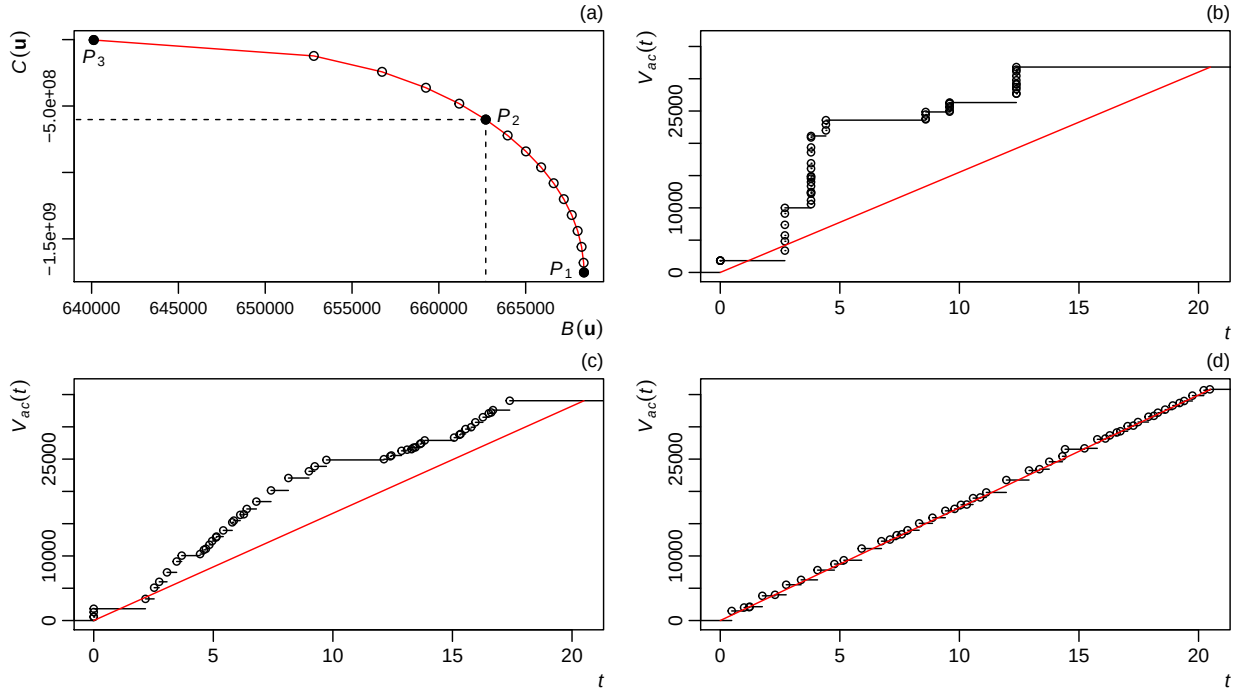


Figura 1: Algunos resultados obtenidos: (a) frente Pareto y soluciones (volumen acumulado obtenido —función escalonada— y volumen acumulado teórico —recta—) correspondientes a los puntos (b) P_1, (c) P_2 y (d) P_3 del frente.

Referencias

[1] Arias-Rodil M., Diéguez-Aranda U., Vázquez-Méndez M.E. (2017). A differentiable optimization model for the management of single-species, even-aged stands. *Canadian Journal of Forest Research* 47 (4), 506–514.

[2] García-Villabrille J.D. (2009). *Modelización del crecimiento y la producción de la plantación de Eucalyptus globulus Labill. en el noroeste de España.* Tesis Doctoral, Universidade de Santiago de Compostela.

[3] Ducheyne E.I., De Wulf R.R., De Baets B. (2004). Single versus multiple objective genetic algorithms for solving the even-flow forest management problem. *Forest Ecology and Management* 201, 259–273.

[4] Johnson K.N., Scheurman H.L. (1977). Techniques for Prescribing Optimal Timber Harvest and Investment Under Different Objectives–Discussion and Synthesis. *Forest Science* 23 (1), 1–32.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

ESTUDO DE CASO CONTROLO: DILEMAS NO CÁLCULO DO TAMANHO AMOSTRAL

Luzia Gonçalves[1]

[1]Unidade de Saúde Pública Internacional e Bioestatística, Instituto de Higiene e Medicina Tropical (IHMT), Universidade Nova de Lisboa, GHTM e CEAUL.

RESUMO

O cálculo do tamanho amostral para estudos de caso-controlo é um assunto antigo. No entanto, na prática nem sempre a teórica se adequa às características específicas de uma investigação médica. Num estudo de caso-controlo clássico, numa relação 1:1, ainda é frequente o cálculo do tamanho amostral com foco em apenas uma exposição de natureza binária. Para os estudos de caso-controlo emparelhados numa relação de 1:M, também existem soluções matemáticas, mas menos acessíveis nos programas usados na prática biomédica. Por vezes, a definição de "caso" original dá posteriormente origem a "caso de tipo 1" e a "caso de tipo 2", originado um estudo duplo de caso-controlo ou estudo de caso-caso-controlo. Neste trabalho, exploram-se alguns "dilemas" nas opções práticas quando o enquadramento teórico disponível não corresponde totalmente às características e objetivos do estudo a implementar.

Palavras e frases chave: Tamanho amostral, estudo caso controlo, estudo caso-caso controlo.

1. INTRODUÇÃO

Por definição um estudo de caso-controlo é um estudo observacional descritivo, frequentemente, de natureza retrospetiva (Keogh e Cox, 2014). Porém, por vezes podemos planear um estudo para o futuro, estudando os casos incidentes e não os casos prevalentes. A definição dos casos deve ser clara e inequívoca. Estes estudos partem dos indivíduos doentes (casos), podendo estes serem ou não emparelhados com os indivíduos que não têm essa doença (controlos), recolhendo informação prévia sobre a presença ou ausência de exposição a determinados fatores que podem ser importantes para o desenvolvimento da doença. A seleção dos controlos, vindos da mesma população que originou os casos, e independente das exposições, é também um aspeto delicado. Por vezes, os controlos são selecionados aleatoriamente da população que originou os casos, outras vezes são controlos hospitalares, outras são familiares ou vizinhos dos casos (e.g., Keogh e Cox, 2014). Os estudos caso-controle geralmente investigam doenças raras, sendo os vieses de seleção e de memória algumas das limitações descritas (e.g., Keogh e Cox, 2014). Estes estudos

permitem estudar simultaneamente várias exposições com possível associação com a doença. Pelo baixo custo, pela rapidez e facilidade de implementação e, ainda, porque o foco continua a centrar-se em exposições que poderão ser eventuais fatores de risco para a doença em causa, podem ser usados também numa situação de surtos epidémicos, ou doenças desconhecidas, em que é essencial agir rapidamente com vista à obtenção de informação que contribua para ações de controlo. Por exemplo, no âmbito do aumento de casos de celulite necrotizante em São Tomé e Príncipe, Gonçalves e Monteverde (2017) analisam um estudo de caso-controlo, com controlos hospitalares e com emparelhamento por sexo, idade e data de internamento. Nesse estudo, a magnitude do *odds ratio* (ou razão das chances) aponta para que indivíduos com lesões/feridas, nos últimos 15 dias, tenham cerca de 5 ou 6 vezes mais chances de terem celulite necrotizante. Para diversas exposições ambientais e ocupacionais não se encontram diferenças significativas entre casos e controlos.

Numa situação de surto ou doenças desconhecidas pode ser difícil ter estimativas iniciais para os parâmetros de forma a calcular o tamanho amostral para os casos e os controlos. Por outro lado, por vezes sabemos que temos *n* casos disponíveis no presente, e aí pode fazer sentido tentar explorar qual será a potência do estudo selecionando os controlos numa relação de 1:1 ou 1: M (com $M \leq 4$), com ou sem emparelhamento. Num estudo de caso-controlo clássico, sem emparelhamento e numa relação de 1:1, o cálculo do tamanho da amostra continua a basear-se frequentemente numa única exposição de natureza binária, mesmo que posteriormente a análise estatística inclua várias exposições numa regressão logística múltipla não condicionada. Demidenko (2008) e Gail e Haneuse (2017) apresentam desenvolvimentos no sentido de alagar o espectro de opções mais compatíveis com a realidade. Num estudo de caso-controlo com emparelhamento de casos e de controlo num rácio de 1 caso para *M* controlos, com vista ao controlo de possíveis efeitos de confundimento (e.g., idade e sexo), quer no cálculo do tamanho da amostra (Gauderman, 2002), quer na análise estatística posterior (Aigner et al, 2018; Pearce, 2016; Niven et al., 2012), existem diversas discussões e lacunas de base que deverão motivar novos desenvolvimentos estatísticos.

Por vezes, os casos podem ser separados em diferentes tipos, por exemplos, casos de tipo 1 e casos de tipo 2, dando origem a um estudo caso-caso-controlo. Para estudar resistencias antimicrobianas, este tipo de estudo tem sido usado (Kaye et al, 2005). Por exemplo, os casos de uma infeção são separados em casos resistentes (casos 1) e casos suscetíveis (casos 2), comparando-os com os respetivos controlos, separadamente, identificando possíveis fatores de risco para ambas as situações.

2. MÉTODOS: ALGUMAS OPÇÕES USANDO ISOGRÁFICOS

A função `epi.ccsize` do pacote epiR do Programa R (Stevenson, 2017, R Development Core Team, 2018) permite o cálculo do tamanho amostral, a potência ou o valor mínimo a ser detetado para o *odds ratio* (OR), quer para estudos de caso-controlo emparelhados, quer para estudos não emparelhados, seguindo as fórmulas propostas por Dupont (1988) e exemplos de Woodward (2005). Através deste pacote estatístico podem-se construir isográficos que poderão ser úteis quando implementamos um estudo baseado em casos incidentes. Assim, por exemplo, estabelecendo diferentes valores para a magnitude do OR e a proporção de exposição no grupo

dos controlos, podemos representar simultaneamente diferentes dimensões das amostras de casos, como se ilustra na Figura 1, para um estudo emparelhado [1:3] com potência de 80%.

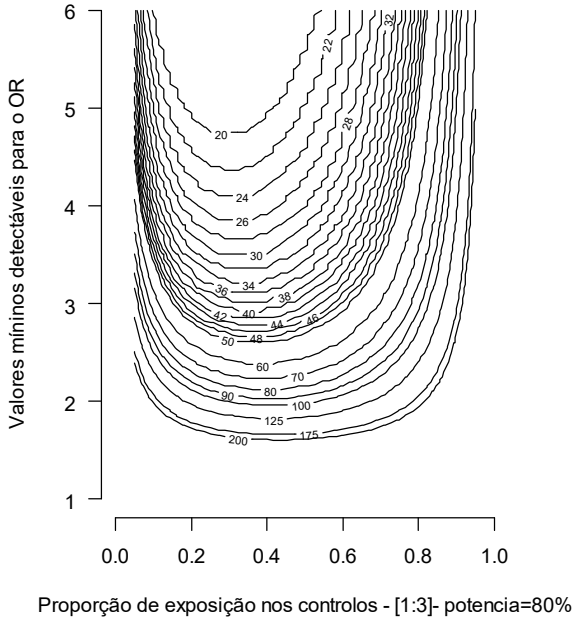


Figura 1. Isográfico com os tamanhos amostrais em função dos valores mínimos detetáveis para o *odds ratio* e diferentes proporções de exposição nos controlos, num estudo de caso-controlo emparelhado numa relação de 1:3, com potência de 80%.

Note-se que muitas vezes ao conhecer o contexto onde se vai realizar o estudo, podemos ter uma noção da dificuldade em recrutar os casos e os controlos. Relativamente ao estudo da celulite necrotizante em de São Tomé e Príncipe, suponhamos que queremos fazer um estudo mais alargado com controlos comunitários, em 7 semanas, e que a escolha recaía em 80 casos e 240 controlos, permitindo detetar OR>2. Além disso, consideremos que os casos vão ser subdivididos em caso de tipo 1 (evolução não complicada) e casos de tipo 2 (situações necrotizantes), após o acompanhamento clínico e laboratorial, respetivamente com frequências p_1 e p_2 (com $p_1 + p_2 = 1$ e $p_1 > p_2$). O número máximo de novos casos (108 casos) foi registado na semana 50 de 2016. Na semana 13 de 2018, registaram 20 casos (WHO Health Emergencies Programme 2017; 2018). Esperando-se uma eventual diminuição do número de casos nessas 7 semanas, mesmo que o número médio por semana seja de 12 casos, que cumpram os critérios de inclusão, é viável recolher o tamanho amostral calculado, desde que o orçamento e outros aspetos logísticos também o permitam. Porém, numa situação de caso-caso-controlo, a amostra dos casos de tipo 2 pode ser reduzida (e.g. se p_2=0.30, n_2=24) e apenas permitir a identificação de associações correspondentes a OR>4.

3. CONCLUSÕES

Mesmo sendo um problema antigo, o enquadramento teórico para algumas variantes dos estudos de caso-controlo ainda requer desenvolvimentos adicionais. Por vezes, o conhecimento da dinâmica do aparecimento dos casos, as questões éticas, o tempo para a investigação e o orçamento, entre outros aspetos, são determinantes, mas a justificação teórica é essencial.

AGRADECIMENTOS

Trabalho financiado por FCT - Fundação para a Ciência e Tecnologia, Portugal (UID/MAT/00006/2013 e UID/Muli/04413/2013).

Referências

[1] Aigner, A., Grittner, U., Becher, H. (2018). Bias due to differential participation in case-control studies and review of available approaches for adjustment. PLoS ONE 13(1): e0191327.

[2] Demidenko, E. (2008) Sample size and optimal design for logistic regression with binary interaction. *Statistics in Medicine*, 27:36-46

[3] Dupont, W. D. (1988). Power Calculations for Matched Case-Control Studies. *Biometrics* 44, 1157–1168.

[4] Gail, M.H., Haneuse, S. (2017). Power and sample size for multivariate logistic modeling of unmatched case-control studies. *Statistical Methods in Medical Research*,

[5] Gauderman, W.J. (2002). Sample size requirements for matched case-control studies of gene–environment interaction. *Statistics in Medicine*, 21:35-50.

[6] Gonçalves, L., Monteverde, E. (2017). *Análise do estudo de caso-controlo e dados complementares sobre a celulite necrotizante*. Ministério da Saúde de São Tomé e Príncipe e Instituto de Higiene e Medicina Tropical Universidade Nova de Lisboa.

[7] Kaye, K.S., Harris, A.D., Samore, M., Carmeli, Y. (2005) The case-case control study design: Addressing the limitations of risk factor studies for antimicrobial resistance. *Infect Control Hosp Epidemiol*,; 26(4):346-51.

[8] Keogh, R.H., Cox, D.R. (2014). *Case-Control Studies*. Cambridge University Press.

[9] Niven, D.J., Berthiaume, L.R., Fick, G.H., Laupland, K.B. (2012). Matched case-control studies: a review of reported statistical methodology. *Clinical Epidemiology* 4(1):99-110.

[10] Pearce, N. (2016). Analysis of matched case-control studies. *BMJ* 352:i969.

[11] R Development Core Team (2018). *R Foundation for Statistical Computing*. Vienna, Austria.

[12] Stevenson, M. (2017). epiR: *Tools for the analysis of epidemiological data*. Available at: ftp://cran.r-project.org/pub/R/web/packages/epiR/epiR.pdf.

[13] WHO Health Emergencies Programme (2017). *Weekly Bulletin on Outbreaks and Other Emergencies*. World Health Organization Africa (Health Emergencies Information and Risk Assessment). Report No.: 33.

[14] WHO Health Emergencies Programme (2018). *Weekly Bulletin on Outbreaks and Other Emergencies*. World Health Organization Africa (Health Emergencies Information and Risk Assessment). Report No.: 13.

[15] Woodward, M. (2005). *Epidemiology Study Design and Data Analysis*. Chapman and Hall/CRC.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

AGREEMENT BETWEEN REGIONAL CLIMATE PROJECTIONS FROM DIFFERENT EURO-CORDEX MODELS: AN EXPLORATORY STUDY

Ana Martins[1], Sandra Rafael[2], Alexandra Monteiro[2], Manuel Scotto[3], Sónia Gouveia[4]

[1] Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, Portugal;
[2] Centre for Environmental and Marine Studies (CESAM), University of Aveiro, Portugal;
[3] CEMAT and Department of Mathematics, IST, University of Lisbon, Portugal;
[4] IEETA and Center for R&D in Mathematics and Applications (CIDMA), University of Aveiro, Portugal;

ABSTRACT

The Euro-Cordex produces regional climate projections based on multiple dynamical and empirical-statistical downscaling models. Nowadays, several climate models are available in Euro-Cordex and, thus, several projections can be obtained for each spatial location. The goal of this work is to analyse the agreement between projections of 2 climate models in 113 meteorological Portuguese stations. The variables temperature, solar radiation, and precipitation were considered for the historical period (1971-2005) and a future scenario (2006-2040). The agreement was measured as the normalised area of the frequency coherence function (averaged over time), being 0 for no agreement and 1 for total agreement between model's projections on a climate variable for a specific location.
Results showed that the model's agreement ranged from 0.34 to 0.62 for all cases (variable, location and historical/future scenario). For all variables, model's agreement exhibited no significant mean differences for the historical and the future periods ($p > 0.05$). Temperature agreement values ranged from 0.44 to 0.62, with a distinct spatial pattern of increasing agreement from south to north of the country. Finally, values ranged between 0.34-0.39 and 0.46-0.57 for solar radiation and precipitation, respectively, and no evident spatial pattern was observed for both variables. In summary, the agreement between models is not consistent for all variables (higher for temperature and lower for solar radiation) and latitudes (for temperature, the agreement is higher in north and lower at lower latitudes).

Keywords and key sentences: Euro-Cordex, Climate projections, agreement method, frequency coherence.

1. INTRODUCTION

According to the recent Intergovernmental Panel on Climate Change (IPCC) 5th Assessment Report, warming of the climate system is unequivocal [**?**]. The assessment of climate change

impacts assessments, for example, to develop local-scale adaptation strategies, requires the availability of high-resolution climate change scenarios. Euro-Cordex (Coordinated Regional Downscaling Experiment, `http://www.euro-cordex.net/`) is the European branch of the international initiative that aims at producing improved regional climate change projections for all regions worldwide. Twenty-six modeling groups contributing with 11 different regional climate models actively support EURO-CORDEX [**?**]. Two regional models, SMHI-RCA4 and CLMCOM-CCLM4-8-17, driven by the global model MPI-M-MPI-ESM-LR (Max Planck Institute Earth System Model) - were selected, focusing on Portugal (mainland) with a spatial resolution of 0.11º. Two climate periods were analysed: historic (1971-2005) and short-term future climate (2006-2040). This global model was selected due to its capability to accurately simulate the European climate [**?**]. As for the regional models SMHI-RCA4 [**?**] and CLMCOM-CCLM4-8-17 [**?**], these differ essentially in terms of model parameterisations, such as convection, microphysics or radiation, among others [**?**].
The main concern in climate projections is the assessment of their robustness and their inherent uncertainties. One of the sources of uncertainty is the variability generated in projections when using different models [**?**]. Thus, the purpose of this study is to explore the agreement of both models in the 113 meteorological Portuguese stations. We considered the temperature, precipitation and solar radiation, for historical and for future climate projections.

2. METHODS

Data from two regional models were used: SMHI-RCA4 and CLMCOM-CCLM4-8-17, both forced by the MPI-M- MPI-ESM-LR global model. A summary of the grid configuration and differences in the parameterisation schemes for the regional models can be found on Jacob *et al.* (2014) [**?**]. Two different time periods were selected for the analysis: historical data, from 1971 to 2005, and a short-term climate projection from 2006 to 2040. The short-time climate projection was based on the Representative Concentration Pathways RCP8.5 scenario, which represents an increase of 8.5 Wm^{-2} in the radiative forcing in the year 2100 comparative to 1750. This is the scenario with the highest increase in radiative forcing. The resolution used for both regional models was 0.11º (≈12.5 km) and the variables analysed were temperature (K), solar radiation (Wm^{-2}) and precipitation (mm).
Data from each model, meteorological variable and time period were downloaded from Euro-Cordex in netCDF format, which contained information for all Europe. Data were preprocessed in *python* in order to extract the information for the 113 Portuguese meteorological stations. The CLMCOM-CCLM4-8-17 model produces daily projections at 12.00 am for all variables, whereas the SMHI-RCA4 model provides measures for every 3 hours (starting at 00:00 for temperature and at 1:30 for solar radiation and for precipitation). Therefore, the time series were resampled for daily observations at 12:00 using spline interpolation ($\Delta t = 1$ day). Note that temperature series represent maximum daily values, whereas solar radiation and precipitation series represent a daily indicator of that variable. Next, the agreement between series $x(t)$ and $y(t)$ was explored by the computation of the magnitude-squared coherence [**?**], defined as

$$C_{xy}(f) = \frac{|G_{xy}(f)|^2}{G_{xx}(f)G_{yy}(f)} \tag{1}$$

where $G_{xy}(f)$ is the cross-spectral density between x and y, and $G_{xx}(f)$ and $G_{yy}(f)$ is the spectral density of x and y, respectively. Note that $C_{xy}(f)$ is defined for frequencies $0 \leq f \leq 1/(2\Delta t)$ and varies between 0 and 1, where 1 indicates a perfect linear relationship between x and y at frequency f. The agreement between x and y was denoted by $\mathcal{A}$ and quantified as the area of $C_{xy}(f)$ over the frequency range divided by $2\Delta t$. This normalisation implies that $0 \leq \mathcal{A} \leq 1$ where $\mathcal{A} = 0$ and $\mathcal{A} = 1$ indicate no agreement and total agreement, respectively. The above mention procedures were applied sequentially, considering $k = 98$ temporal blocks

of 256 observations with 50% of data overlap. This approach provided a sequence of k $C_{xy}(f)$ functions and $\mathcal{A}$ values, one for each block. For each comparison, the agreement profile and overall agreement were obtained as the averaged $C_{xy}(f)$ and averaged $\mathcal{A}$ over blocks. The methods were implemented in R software (version 3.4.2) using the packages *seewave* and *stats*.

3. RESULTS

There is a large overlap between the range of agreement values when comparing historical against RCP8.5 scenario for all variables (result not shown). Also, no significant differences were observed between the mean agreement of historical and RCP8.5 scenario ($p > 0.05$), thus suggesting a similar agreement between the models on past and future projections. Figure 1 shows the mean agreement between the models for the historical period. Temperature agreement ranges from 0.44 to 0.62. Furthermore, there is a clear spatial pattern suggesting that model agreement depends on latitude, being higher for meteorological stations in the north of the country. With respect to precipitation, the agreement is similar to that of temperature (range 0.47-0.56) and no spatial pattern was evident. Finally, solar radiation agreement is similar throughout the country and is fairly low (range 0.34-0.39) for both historical and RCP8.5 periods. This low agreement between models on solar radiation projections can be attributed to model differences in cloud behavior [**?**].

(a) Temperature (b) Solar radiation (c) Precipitation

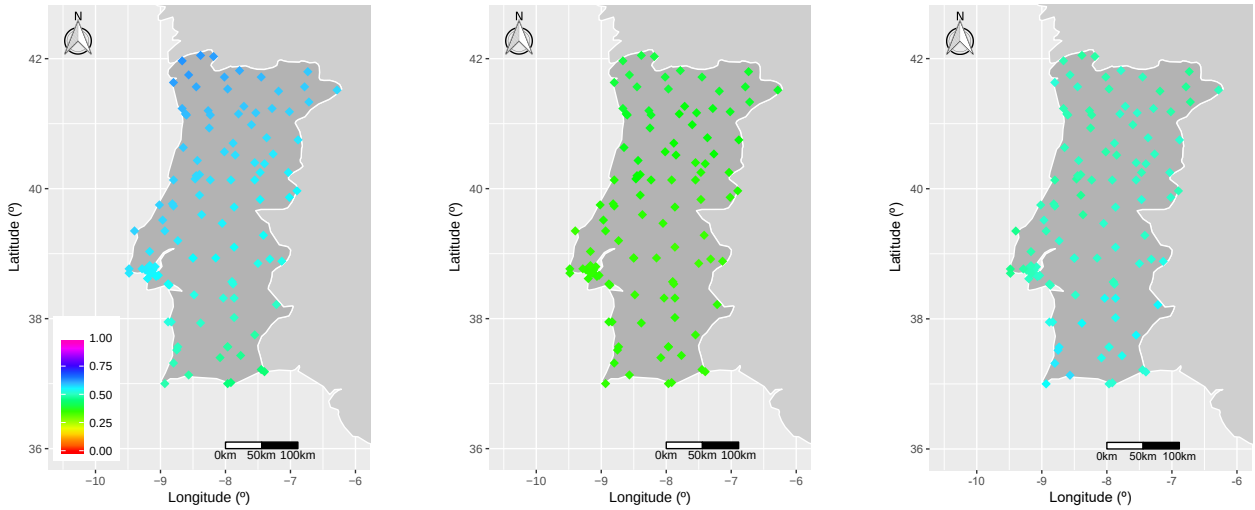


Figure 1: Overall agreement of the CLMCOM-CCLM4-8-17 and SMHI-RCA4 models on historical projections for temperature, solar radiation and precipitation. The dots locate the 113 stations in mainland Portugal and the color refers to the agreement value in each station.

Temperature exhibits a spatial variation that was further investigated. Figure 2 shows the agreement profile for Viana do Castelo, Aveiro and Faro stations on temperature historical data. The mean agreement profile for lower frequencies (up to 0.1 days^{-1}, which correspond to time periods longer than 10 days) is above 0.5 for these stations. This result was expected since as values are averaged over an extense time period. For higher frequencies (and shorter time periods), the agreement decreases rapidly for Faro whereas the agreement is still above to 0.5 for Viana do Castelo and Aveiro, up to frequencies lower than 0.4 days^{-1}. Finally, figure 2 also illustrates that the width of the variability bands is fairly similar for all stations (inter-station comparison) and that the width is approximately constant for all frequencies (intra-station comparison). These results illustrate that the variance of the agreement values is similar for all frequencies and does not depend on the location of the meteorological station.

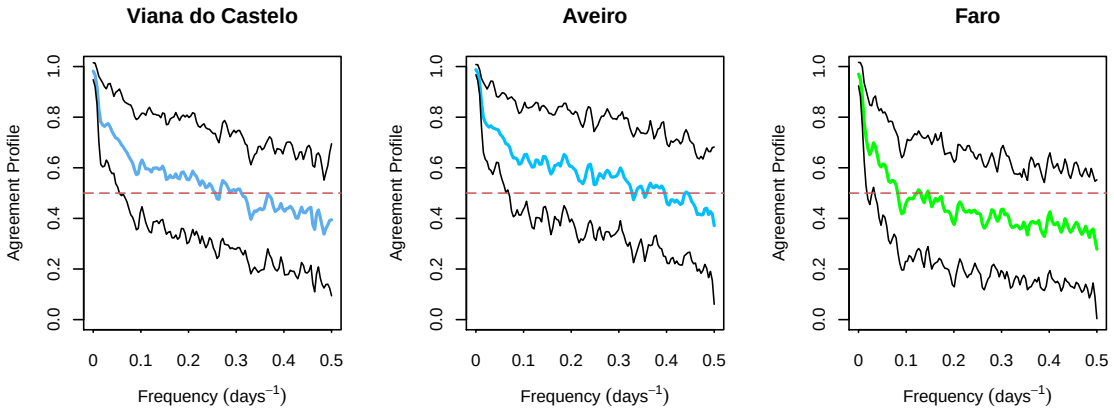


Figure 2: Agreement profile for Viana do Castelo, Aveiro and Faro (temperature historical data). Full lines represent mean ± standard deviation over temporal blocks, and mean is colored according to legend in Figure 1. The dashed line locates the agreement value of 0.5.

4. CONCLUSIONS

In conclusion, the model's agreement was similar for historical and future RCP8.5 scenarios. Overall agreement values ranged between 0.34 and 0.62 for all cases, indicating that model agreement is not particularly high for any of the climate variables in the Portuguese stations. Furthermore, a distinct pattern of increasing agreement from south to north of the country was found for temperature, suggesting that the model's agreement depends on latitude.

Acknowledgements

This work was partially funded by the Foundation for Science and Technology (FCT), through national funds (MEC) and european structural (FEDER), through the UID/CEC/00127/2013 (IEETA) and UID/MAT/04106/2013 (CIDMA) projects. Ana Martins acknowledges the R&D grant in the scope of IEETA project.

References

[1] Stocker, T. F., et al. (2013). Fifth assessment report of the intergovernmental panel on climate change. *The Physical Science Basis*.

[2] Marta-Almeida, M., Teixeira, J. C., Carvalho, M. J., Melo-Gonçalves, P., & Rocha, A. M. (2016). High resolution WRF climatic simulations for the Iberian Peninsula: model validation. *Physics and Chemistry of the Earth*, Parts A/B/C, 94, 94-105.

[3] Samuelsson, P., Jones, C. G., Willén, U., Ullerstig, A., Gollvik, S., Hansson, U. L. F., & Wyser, K. (2011). The Rossby Centre Regional Climate model RCA3: model description and performance. *Tellus A*, 63(1), 4-23.

[4] Rockel, B., Will, A., & Hense, A. (2008). The regional climate model COSMO-CLM (CCLM). *Meteorologische Zeitschrift*, 17(4), 347-348.

[5] Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O. B., Bouwer, L. M., & Georgopoulou, E. (2014). EURO-CORDEX: new high-resolution climate change projections for European impact research. *Regional environmental change*, 14(2), 563-578

[6] Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, 90(8), 1095-1107.

[7] Kay, S. M. (1988) Modern Spectral Estimation. Englewood Cliffs, NJ: Prentice-Hall.

[8] Lobell, D. B., Bonfils, C., & Duffy, P. B. (2007). Climate change uncertainty for daily minimum and maximum temperatures: a model inter-comparison. *Geophysical Research Letters*, 34(5).

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

A ESCOLHA ESTATÍSTICA E A ESTIMAÇÃO EM TEORIA DE VALORES EXTREMOS: APLICAÇÃO EM DADOS AMBIENTAIS

Manuela Neves[1], Helena Penalva[2], Sandra Nunes[3] e Dora Prata Gomes[4]

[1]Instituto Superior de Agronomia, e CEAUL, Universidade de Lisboa, Portugal
[2]Escola de Ciências Empresariais do Instituto Politécnico de Setúbal, e CEAUL, Universidade de Lisboa, Portugal
[3]Escola de Ciências Empresariais do Instituto Politécnico de Setúbal, e CMA/FCT, Universidade Nova de Lisboa, Portugal
[4]Faculdade de Ciências e Tecnologia e CMA/FCT, Universidade Nova de Lisboa, Portugal

RESUMO

A Teoria de Valores Extremos tem vindo a afirmar-se como uma das mais importantes teorias estatísticas para as ciências aplicadas, fornecendo uma base teórica sólida para a construção de modelos estatísticos descrevendo eventos extremos. A inferência e a estimação de parâmetros é baseada nas observações que, nalgum sentido, são consideradas extremas. A eficiência dos procedimentos adoptados depende da forma da cauda da distribuição subjacente aos dados. Neste trabalho iremos apresentar uma revisão de testes à cauda da distribuição e proceder depois à estimação mais adequada de parâmetros de interesse. Os procedimentos são aplicados a conjuntos de dados ambientais.

Palavras e frases chave: Cauda da distribuição; escolha estatística; estimação semi-paramétrica; teoria de valores extremos; testes estatísticos.

1. INTRODUÇÃO

A Teoria de Valores Extremos (usualmente denotada por EVT, do inglês *Extreme Value Theory*) tem como objectivo o estudo de fenómenos nos quais poderão ocorrer valores muito para além do que conseguimos observar – a essas ocorrências chama-se, por isso, *acontecimentos raros*. As primeiras aplicações da EVT surgiram com a modelação de fenómenos meteorológicos envolvendo precipitações máximas e inundações. Contudo, a abrangência das suas aplicações é muito vasta, incluindo uma variedade de fenómenos naturais tais como poluição atmosférica, correntes oceânicas, rajadas de vento, sismos e problemas oriundos de outras áreas tais como da engenharia, actuariado e finanças. A modelação e a inferência em teoria de valores extremos é da maior relevância em situações em que possam ocorrer catástrofes.
A interpretação de dados de valores extremos com a escolha adequada da cauda representa um desafio importante quando se lida com aplicações práticas reais. A escolha do tipo de cauda da distribuição subjacente aos dados, deve constituir uma análise prévia à estimação

dos parâmetros de interesse. Os testes de ajustamento usados tradicionalmente na inferência estatística, não fornecem informação adequada sobre a forma das caudas, tendo por isso surgido testes adequados a valores extremos.
Sendo $(X_1, X_2, \cdots, X_n)$ uma amostra de variáveis aleatórias independentes e identicamente distribuídas ou possivelmente fracamente dependentes, Fréchet (1927), Fisher and Tippet (1928), Gumbel (1935) e von Mises (1936) apresentaram os primeiros resultados do chamado *problema do limite extremal*. Porém foram Gnedenko (1943) e de Haan (1970) que formularam as condições para a existência de sucessões $\{a_n\} \in R^+$ e $\{b_n\} \in R$ tais que,

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = EV_\xi(x) \quad \forall x \in R, \tag{1}$$

onde EV_ξ é uma função de distribuição não degenerada. Esta função, designada função distribuição de *Valores Extremos* é dada por

$$EV_\xi(x) = \begin{cases} \exp\left\{-\left[1+\xi x\right]^{-1/\xi}\right\}, & 1+\xi x > 0, \ se\ \xi \neq 0\ ; \\ \exp\left\{-\exp\left[-x\right]\right\}, & x \in R, \ se\ \xi = 0, \end{cases} \tag{2}$$

onde $\xi \in R$ é o parâmetro de forma, habitualmente designado por *índice de valores extremos*. Quando (1) se verifica dizemos que F está no domínio de atracção (para máximos) de EV_ξ e escrevemos $F \in \mathcal{D}_\mathcal{M}(EV_\xi)$.
O parâmetro de forma, ξ, mede o peso da cauda direita, $\overline{F} := 1-F$, da distribuição subjacente. Se $\xi = 0$, a cauda direita é do tipo exponencial. Se $\xi > 0$, a cauda direita é pesada, é do tipo polinomial negativo e se $\xi < 0$, a cauda direita é curta e F tem um *limite superior de suporte* finito.
Em EVT a análise estatística e a inferência podem ser realizadas numa abordagen paramétrica ou numa aborgadem semi-paramétrica. Em qualquer uma destas abordagens presume-se que a função de distribuição subjacente F pertence a $\mathcal{D}_\mathcal{M}(EV_\xi)$, para um valor apropriado de ξ, ou está num subdomínio específico de $\mathcal{D}_\mathcal{M}(EV_\xi)$. Essa condição é chamada de *condição de valores extremos* e, face a um conjunto de dados, é importante ser verificada, i.e. deve testar-se:

$$H_0 : F \in \mathcal{D}_\mathcal{M}(EV_\xi) \text{ para algum } \xi \in R.$$

Podemos referir alguns testes para aquela hipótese como os estudados em Dietrich et al.(2002), Drees et al. (2006) e Hüsler and Li (2006).
Além destes testes faremos também referência ao problema da escolha entre um dos três tipos extremais, dando preferência ao modelo de Gumbel para a hipótese nula; este problema recebeu a designação geral de *escolha estatística dos modelos extremais*, foi tratado por vários autores e faremos adiante uma ilustração com um dos conjuntos de dados considerado.
O objectivo deste trabalho é apresentar uma breve revisão geral de vários testes para a escolha estatística da cauda, para permitir uma estimação mais adequada dos parâmetros de interesse. Dos parâmetros de acontecimentos raros cuja estimação tem relevância especial podemos indicar: o índice de valores extremos, ξ; quantis elevados, i.e, valores χ_p, com p pequeno; o período de retorno de um nível elevado u , i.e, tempo médio de espera entre excedências independentes desse nível; o limite superior do suporte de uma distribuição F, w_F, e o índice extremal, θ, parâmetro que desempenha papel fundamental na passagem de uma estrutura independente para uma dependente e que, informalmente, pode ser definido como o recíproco da duração média de acontecimentos extremos.

2. UM EXEMPLO DE APLICAÇÂO

Iremos então fazer aplicação de testes à condição de valores extremos, testes à cauda e ainda estimação de parâmetros de interesse considerando dois exemplos: o primeiro conjunto de dados refere-se aos valores diários de área ardida em Portugal Continental no período 1990–2003 e o segundo conjunto de dados é referente à velocidade máxima diária de vento (m/s) registada em Lisboa de Outubro de 1946 a Setembro de 2006, recolhida no SNIRH: Sistema Nacional de Informação dos Recursos Hídricos.
Como ilustração apresentamos na seguinte figura as trajectórias amostrais referentes aos testes à cauda para o primeiro conjunto de dados. As estatísticas de testes aqui ilustradas, e que se encontram descritas e estudadas em pormenor em Neves e Fraga Alves (2007, 2008), são designadas por G^*, R^* e W^*.
Para aquele conjunto de dados foi realizado o teste

$$H_0 : F \in \mathcal{D}_{\mathcal{M}}(EV_0) \quad vs \quad H_1 : F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi>0}.$$

vendo-se as trajectórias amostrais de G^*, R^* e W^* para alguns valores de k.
As três estatísticas de teste apresentam valores que pertencem às correspondentes regiões de rejeição podendo-se assim concluir pelo domínio de atracção da Fréchet.
Procedeu-se de seguida à escolha de estimadores adequados para a estimação de alguns parâmetros.

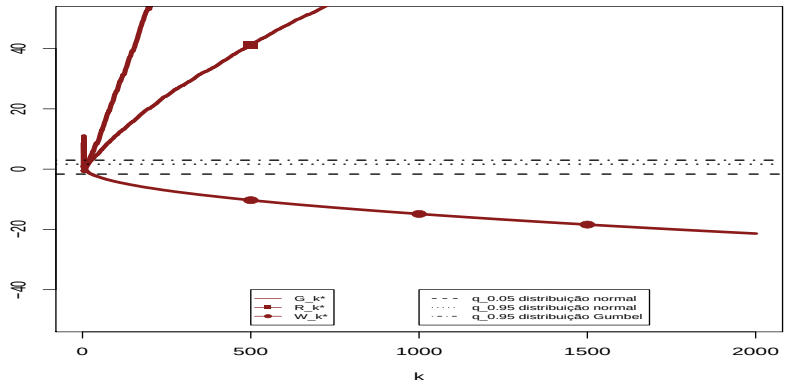


Figura 1: Trajectórias amostrais de G^*, R^* e W^* .

Em Neves et. al (2015) encontramos outro exemplo de aplicação.

AGRADECIMENTOS

Investigação parcialmente financiada pelos Fundos Nacionais da FCT–Fundação para a Ciência e a Tecnologia, projectos UID/MAT/00006/2013 (CEAUL) e PEst-OE/MAT/UI0297/2013 (CMA/UNL).

Referências

[1] Dietrich, D., de Haan, L., and Husler, J. (2002). Testing extreme value conditions. Extremes, 5, 71–85.

[2] Drees, H., de Haan, L., and Li, D. (2006). Approximations to the tail empirical distribution function with application to testing extreme value conditions. Journal of Statistical Planning and Inference, 136, 3498–3538.

[3] Fisher, R. A. and Tippett, L. H. C. (1928). On the estimation of the frequency distributions of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society, 24, 180-190.

[4] Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. Ann. Soc. Polon. Math. (Cracovie), 6, 93-116.

[5] Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d'une série aléaatoire. Annals of Mathematics, 44, 423–453.

[6] Gumbel, E. J. (1935). Les valeurs extrêmes des distributions statistiques. Ann. Inst. Henri Poincaré, 5(2), 115-158.

[7] de Haan, L. (1970). On Regular Variation and its Applications to the Weak Convergence of Sample Extremes, Mathematical Centre Tract 32, Amesterdam, Dordrecht: D.Reidel.

[8] Husler, J. and Li, D. (2006). On testing extreme value conditions. Extremes, 9, 69–86.

[9] Neves, C. and Fraga Alves, M. I. (2007). Semi-parametric approach to Hasofer-Wang and Greenwood statistics in extremes. Test, 16, 297-313.

[10] Neves, C. and Fraga Alves, M. I. (2008). Testing extreme value conditions– an overview and recent approaches. Revstat, 6(1), 83–100.

[11] Neves, M.M., Penalva, H. and Nunes, S. (2015). Extreme value analysis of river levels in a hydrometric station in the North of Portugal. Current Topics on Risk Analysis: ICRA6 and RISK 2015 Conference Proceedings. M. Guillén, A. Juan, H. Ramalhinho, I. Serra and C. Serrat (edts), 533–538.

[12] von Mises, R. (1936). La distribution de la plus grande de n valeurs. Rev. Math. Union Interbalcanique, 1:141–160. Reprinted in Selected Papers of Richard von Mises, Amer. Math. Soc.(1964), 2:271–294.

III Encontro Luso-Galaico de Biometría
Aveiro, 28 - 30 June 2018

EFFECTS OF A HEALTH EDUCATION INTERVENTION ON PHYSICAL ACTIVITY IN INDIVIDUALS WITH MODERATE-TO-HIGH CARDIOVASCULAR RISK

Lucimere Bohn[1,2], Pedro Sa-Couto[3], Ana Ramoa Castro[4], Fernando Ribeiro[5], José Oliveira[1]

[1] Faculty of Sport, University of Porto. Research Centre in Physical Activity, Health and Leisure, University of Porto, Portugal
[2] Escola Superior de Desporto e Lazer. Instituto Politécnico de Viana do Castelo, Portugal
[3] Center for Research and Development in Mathematics and Applications and Department of Mathematics, University of Aveiro, Aveiro, Portugal.
[4] Primary Care Centre Espaço Saúde. Aldoar, Porto, Portugal
[5] School of Health Sciences and Institute of Biomedicine–iBiMED, University of Aveiro, Aveiro, Portugal

ABSTRACT

This study evaluated the effects of a health education and counseling intervention program, in a primary healthcare setting, on daily physical activity (PA) in individuals with moderate-to-high risk of cardiovascular disease. This was a parallel-group study with a 4-month-long intervention, plus 8 months of follow-up. Participants were 164 individuals with moderate-to-high cardiovascular risk, allocated to either an intervention (IC, n=87) or a control group (CG, n=77). The intervention consisted by 3 walking and face-to-face group sessions plus text messages. The primary outcome was daily PA measured by sedentary time, light and moderate-to-vigorous PA. After the intervention (4 months) and follow-up (8 months) periods, the results show significant differences between Groups (IC, CG) for sedentary time and light PA, but not for moderate-to-vigorous PA. No significant changes were found for the variable Time (baseline, 4 months, 8 months) and for the correspondent interaction between Groups and Time, even after adjustments for age, gender, BMI, and variables that were different between groups at baseline. The health education and counseling program did not improve daily PA of participants with moderate-to-high cardiovascular risk.

Keywords and key sentences: education, counseling, primary care, cardiovascular risk, daily physical activity, linear mixed models regression

1. INTRODUCTION

Physical activity (PA) confers health benefits, with evidence indicating that any amount of PA is healthful. The increment of daily PA levels is recommended in primary and secondary prevention of cardiovascular disease (CVD). Despite the recommendations, 31.1% of the adults worldwide fail to meet the PA guidelines.[1]

Given that the incidence of CVD remains high, the early detection of patients at risk is an important strategy to prevent the onset of CVD. In developed countries, 70-80% of adults visit their general practitioner at least once a year, which makes the primary care health services the best setting to assess cardiovascular risk, manage risk factors, and promote a healthy lifestyle, including the promotion of PA. The aim of this study was to evaluate the effects of a 4-month health education and counseling intervention in primary care, and an 8-month follow-up period, on daily PA in adults with moderate to high cardiovascular risk.

2. METHODS

This study was a parallel-group with a non-probabilistic sample conduced from March 2012 to July 2013 at the primary health care center. The study consisted of a health education and counseling intervention aiming to promote the increase in daily PA levels. Participants were selected from the registries of a primary health care center. Allocation to the intervention group (IG) was made by convenience according to the will and availability to participate in educational and counseling group sessions and to receive text messages on their mobile phones. Those who agreed to participate in the evaluations but were not available to participate in the health education and counseling IG were allocated to the control group (CG). The study was approved by the Ethics Committee of the North Regional Health Authority (I.P. 25/2010) and all procedures were conducted according to the Helsinki declaration.

Daily PA was assessed using accelerometers (Actigraph GT1M, Actigraph LLC, Pensacola, FL) over the right hip, for 7 consecutive days, during the waking hours, except while bathing and water-based activities. The average minutes/day spent at different categories of PA intensity was determined according to cut points that relate PA to counts/min: sedentary time (≤ 99 counts/min), light PA (100 - 2019 counts/min) and moderate-to-vigorous MVPA (MVPA) (≥ 2020 counts/min). Total cardiovascular risk was calculated according to the 2013 ESH/ESC International Guidelines for the Management of Hypertension.[2]

The intervention consisted of three group sessions, followed by mobile text messages to encourage and reinforce PA adherence. Two general practitioners and a PA specialist delivered the health educational and counseling program, which was consisted of three sessions, lasting approximately 90 minutes each. A maximum of 10 participants were included in each session group. Sessions were composed of a 30-minute group walk at moderate intensity in the city park, followed by 60 minutes of face-to-face intervention. In the first 60-minute session, a general practitioner presented information about the CVD risk concept, how to identify personal risk factors that influence the CVD risk, and insights about the impact of a moderate and high CVD risk on health status and quality of life. A general practitioner conducted the second session and the content was targeted at healthy behaviors and lifestyle (i.e., diet; tobacco cessation; salt intake; adherence and compliance with medication; stress management; and PA) as a path to diminish CVD risk. The third session was conduced by a PA specialist, the participants received a booklet with all the information presented during the sessions and a PA plans for each week of the four-month period. After the sessions, participants in the IG received 12 mobile text messages to encourage and reinforce PA adherence. The texts messages were

delivered once a week during the first two months, and twice a month in the last 2 months. During the follow-up IG and CG only received the usual care. The intervention program followed the recommendations and standards of the American College of Sports Medicine.[3]

An intention-to-treat analysis was conducted, with the inclusion of all participants assessed and allocated into groups at baseline. Changes in groups (IC; CG), time (baseline, 4 months; 8 months), and groups over time (group*time interaction) were modeled using a linear mixed-model regression with random-effects. The covariance type used for the random-effects was the unstructured option (completely general covariance matrix). Other covariance types (e.g. first order autoregressive) were also used but presented less accurate results (higher Akaike's Information Criterion values). Normality of residuals was verified by visually inspection. Statistical analysis was performed using IBM SPSS software version 21 (SPSS, Chicago, USA) and vales of *P* less than 0.05 were considered significant.

3. RESULTS

The study included 85 participants in the IG (57.16 ± 6.61 years old; males 45.9%) and 77 in the CG (55.42 ± 7.34 years old; males 44.2%) with moderate-to-high CVD risk. Considering daily PA, the IG showed significantly higher sedentary time (p=0.040) and lower light PA (p=0.004) than the CG (Table 1). Also, de body mass index (BMI) has a slightly decrease throughout time (p=0.038) for the IC group.

		Baseline (T1)		4 months (T2)		Follow up (T3)	
		n	**Mean ± sd**	**n**	**Mean ± sd**	**n**	**Mean ± sd**
Body mass index, kg/m^2	IG	85	29.27 (3.91)	76	28.87 (3.91)	56	28.92 (3.79)
	CG	77	29.89 (4.32)	65	30.19 (4.08)	42	28.96 (3.64)
Sedentary time, min/day*	IG	82	472.2 (85.6)	71	452.2 (89.3)	54	454.4 (102.2)
	CG	74	435.4 (100.2)	63	426.2 (109.2)	37	441.3 (101.8)
Light PA, min/day*	IG	82	289.8 (92.4)	71	299.1 (94.4)	54	294.4 (92.5)
	CG	74	337.0 (103.7)	63	331.2 (102.1)	37	320.1 (86.1)
MVPA, min/day	IG	82	32.9 (25.8)	71	41.0 (29.9)	54	34.3 (27.4)
	CG	74	38.3 (31.4)	63	41.0 (30.5)	37	42.5 (39.3)

Table 1. Parameters at baseline, 4 months and 8 months (follow-up) for an intention-to-treat analysis. *Groups were significantly different at baseline p < 0.05

	Factors	**Unadjusted model**	**Adjusted model**
Sedentary time, min/day	Group	41.4 (17.9); p=0.021	42.4 (18.2); p=0.021
	Group * Time	-5.99 (7.6); p=0.433	-7.4 (7.6); p=0.331
Light PA, min/day	Group	-50.0 (18.9); p=0.009	-50.5(19.0); p=0.009
	Group * Time	4.7 (6.5); p=0.468	4.4 (6.4); p=0.491
MVPA, min/day	Group	-2.8 (5.9); p=0.640	-2.7 (5.9); p=0.651
	Group * Time	-0.4 (2.5); p=0.878	0.1 (2.6); p=0.938

Table 2. Linear mixed model regression for sedentary time, light and moderate-to-vigorous physical activity. Values presented are in Slope(SE). The factor time was always non-significant. Adjusted model: age, gender, body mass index, dyslipidemia, antihypertensive and antidepressants/ anxiolytic medication.

After the intervention (4 months) and follow-up (8 months) periods, the results show significant differences between Groups (IC versus CG) for sedentary time and light PA, but not for moderate-to-vigorous PA (Table 2). No significant changes were found for the variable Time (baseline, 4 months, 8 months) and for the correspondent interaction between Groups and Time (Table 2, unadjusted models). After adjustments for age, gender, BMI, and variables that were different between groups at baseline, the results remained similar (Table 2, adjusted models).

4. CONCLUSIONS

In conclusion, this study did not provide evidence for the efficacy of a health education and counseling program in a primary care setting to improve daily PA levels in individuals with moderate to high cardiovascular risk. This study used an objective measurement for PA, which likely improved the accuracy of assessments over time. Indeed, the use of self-report measures of PA is the most common method in previously published trials, which might inflate estimates of interventions effects, once respondents tend to report less sedentary behaviors and more MVPA.[4]

Several limitations of this study should be noted. First, the allocation of patients into groups was made by convenience. Second, this study did not assess self-regulation for PA and compared this between the groups. Given that allocation was made by convenience, it is possible that those included in the IG were those who were more conscious of and motivated bout the importance of lifestyle changes. Third, the sample size, and the participant's retention at one year, was small.

ACKNOWLEDGMENT

The European Regional Development Fund through the Operational Competitiveness Program, and the Foundation for Science and Technology (FCT) of Portugal support this study and the research unit CIAFEL within the projects FCOMP-01-0124-FEDER-020180 (References FCT: PTDC/DES/122763/2010,UID/DTP/00617/2013,UID/BIM/04501/2013,UID/MAT/04106/2013). The FCT supported the author Lucimére Bohn (SFRH/BD/78620/2011).

References

[1] Hallal PC, Andersen LB, Bull FC, Guthold R, Haskell W, Ekelund U (2012). Global physical activity levels: surveillance progress, pitfalls, and prospects. Lancet.;380(9838):247-257

[2] Mancia G, Fagard R, Narkiewicz K, et al. (2013). ESH/ESC Practice Guidelines for the Management of Arterial Hypertension. Blood Press;23(1):3-16.

[3] Garber CE, Blissmer B, Deschenes MR, et al. (2011). American College of Sports Medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise. Med. Sci. Sports Exerc;43(7):1334-1359

[4] Orrow G, Kinmonth AL, Sanderson S, Sutton S (2012). Effectiveness of physical activity promotion based in primary care: systematic review and meta-analysis of randomized controlled trials. BMJ; 344:e1389

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

NONLINEAR MIXED-EFFECTS MODEL FOR CYCLOSPORINE PHARMACOKINETICS IN RENAL TRANSPLANT

A. Sofia Cardoso1[1], M. Salomé Cabral[2] A. Paula Carrondo[3] e José Guerra[4]

[1]Serviço de Gestão Técnico-Farmacêutica, CHLN-Hospital de Santa Maria, Lisboa, Portugal
[2]CEAUL, Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa, Portugal
[3]Serviço de Gestão Técnico-Farmacêutica, CHLN-Hospital de Santa Maria, Lisboa, Portugal, Ciências Farmacológicas, Faculdade de Farmácia, Universidade de Lisboa, Portugal
[4]Serviço de Nefrologia e Transplantação Renal, CHLN-Hospital de Santa Maria, Lisboa, Portugal

ABSTRACT

The aim of this work was to develop a population pharmacokinetic (PK) model of cyclosporine (CsA), an immunosuppressive drug used in kidney transplant. Nonlinear mixed-effects models and Bayesian approaches were used and the final model was validated by internal and external methods. The PK model showed a suitable predictability and can be applied to CsA dosing optimization in late stage renal transplant recipients by using conventional monitoring data. Clinical model evaluation and further research in early stage of transplant are still required.

Keywords and key sentences: cyclosporine, population pharmacokinetics, nonlinear mixed-effects model, Bayesian forecasting, renal transplant.

1. INTRODUCTION

Cyclosporine is an immunosuppressive drug used in renal transplantation with a narrow therapeutic window and a wide inter- and intra-individual PK variability [**?**]. Thus, therapeutic drug monitoring is essential to individualize dosing, to prevent adverse effects and maximize efficacy. Information needed for dosage individualization can be obtained through PK modelling.
Many population PK models of CsA have been published. These models relate relevant covariates with the PK parameters, but the majority uses two- or three-compartment PK models.
The aim of this work was to develop a simpler one-compartment PK model easily applied in a clinical setting, using a late renal transplant Portuguese population. The nonlinear mixed-effects model approach was used and the dose individualization prediction was based in Bayesian forecasting through the best linear predictors (BLUPs)[**?**].

2. MATERIAL and METHODS

Routine monitoring data were retrospectively collected between 2001 and 2009 from 104 renal transplant patients followed in the Transplantation Unit of the Santa Maria Hospital (study approved by the ethics commission, on April 29, 2016 code number 660/15). All patients received oral CsA (Novartis Farma, Portugal) soft capsules twice a day. A total of 682 CsA concentrations taken at steady-state were obtained. Blood samples were collected at pre-dose and at the following post-dose time: 1, 2, 3 and 4h, and analysed on AxSym analyzer (Abbott Laboratories, Abbott Park, IL). Demographic, clinical and laboratory data were also collected: gender, age, weight, height, body surface area (BSA), body mass index (BMI), serum creatinine (cr), creatinine clearance (Clcr), and time after transplant (posTx). The database was randomly split into two groups: the modelling group (n=82 patients, 543 concentrations) used to build the model and the validation group (n=22 patients, 139 concentrations) used for external validation. The one-compartment PK model given by (**??**) was used [**?**, **?**]

$$C_t = \frac{FDKa}{\left(Ka - \frac{Cl}{Vd}\right) \times Vd} \times \left[\left(\frac{1}{1 - e^{-\frac{Cl}{Vd}\tau}} \right) \times e^{-\frac{Cl}{Vd}t} - \left(\frac{1}{1 - e^{-Ka\tau}} \right) \times e^{-Kat} \right] + \epsilon, \qquad (1)$$

where C_t is the concentration of drug at time t, after some dose D administration in some dosing interval of time τ and ϵ is the random error with the usual assumptions. The PK parameters are: Ka (absorption constant rate), Cl/F (apparent clearance), and Vd/F (apparent volume of distribution), where F is the unknown bioavailability.
To avoid numerical problems, the PK parameters were reparametrized as $lV = log(Vd)$, $lCl = log(Cl)$ and $lKa = log(Ka)$ [**?**] and Ka was set at 1.28 h^{-1} ($log(Ka) = 0.247$) [**?**, **?**]. Continuous covariates were also reparametrized using logarithmic transformation after scaled by the mean or by the conventional value. The discrete covariate gender was codified as sex=0 for females (F) and sex=1 for males (M).
The stepwise model building strategy described in [**?**] was used. First, the inter-individual variability of the PK parameters was modelled, in the reparametrized scale, as $lV = \beta_1 + b_{1i}$ and $lCl = \beta_2 + b_{2i}$ for the i^{th} patient, followed by covariate screening, according to a linear model, to explain the random-effects variation.
The residual plot analysis was performed following [**?**]. The stability and performance of the final model were assessed by internal and external validation [**?**]. Internal validation was performed using data splitting and jack-knife methods. To assess the external validation, the final model was applied to the 22 additional patients of the validation group and the individual random effects were estimated using Bayesian forecasting [**?**]. Prediction performance was assessed in terms of bias and precision by calculating the mean prediction error (MSE) and the root mean squared error ($RMSE$), respectively [**?**].
Population PK analysis was performed using the `nlme` function in `nlme` library of `S-Plus` 6 and the `quinModel` function in the same library was used to implement the one-compartment PK model. All the statistical inference was made at a significance level of 5%.

3. RESULTS

The final one-compartment PK population model, had random effects on both parameters (lCl/F and lV/F) with a diagonal variance-covariance matrix. The covariates BSA, age and sex had significantly influenced the lCl/F parameter, and none covariate significantly influenced the lV/F parameter. Table 1 displays the final model parameter estimates where $\beta_i (i = 1, \cdots, 5)$ are the coefficients of the PK parameters in the reparametrized scale.

Parameter	Estimate	SE	t-value	$p - value$
β_1	4.3925	0.0520	84.4602	< 0.0001
β_2	3.3925	0.0594	57.1451	< 0.0001
β_3	0.9284	0.3373	2.7523	0.0062
β_4	-0.2680	0.1133	-2.3661	0.0184
β_5	0.1575	0.0707	2.2272	0.0264

Table 1: Parameters estimates, standard errors, t-value and p-value for the final PK model.

In the original scale the estimate of the population model, based on Table 1, was $Vd/F(l) = 80.8$ and $Cl/F(l/h) = 29.7 \times \left(\frac{BSA}{1.73}\right)^{0.928} \times \left(\frac{age}{50}\right)^{-0.268} \times (1.17)^{sex}$, respectively. Thus, to a typical female patient with 1.73 m^2 of body surface area and 50 years of age, Vd/F and Cl/F were 80.8 l and 29.7 l/h, respectively. Cl/F was 25% higher in a typical patient of 22 years old and 9% lower in a typical patient of 71 years old, compared to a patient of 50 years old. The inter-individual variability of Vd/F and Cl/F were 22.6% and 40.7%, respectively, suggesting that Vd/F inter-individual variability may not be fully explained. The standard deviation of the random error was 137.2 ng/ml. The standardized residuals (Figure **??**) were acceptable either in homoscedasticity or in normality conditions, although low pre-dose concentrations were slightly shifted. Figure **??** displays individual and population predictions for two patients showing a good agreement with observed concentrations. Internal validation showed that the PK parameters from the full dataset lied all in the 95% confidence interval of the parameters estimated by data splitting and jack-knife. In the external validation, MPE was 31.05 ng/ml, significantly different from zero [$IC95\%$: $(10.8, 51.39)$].

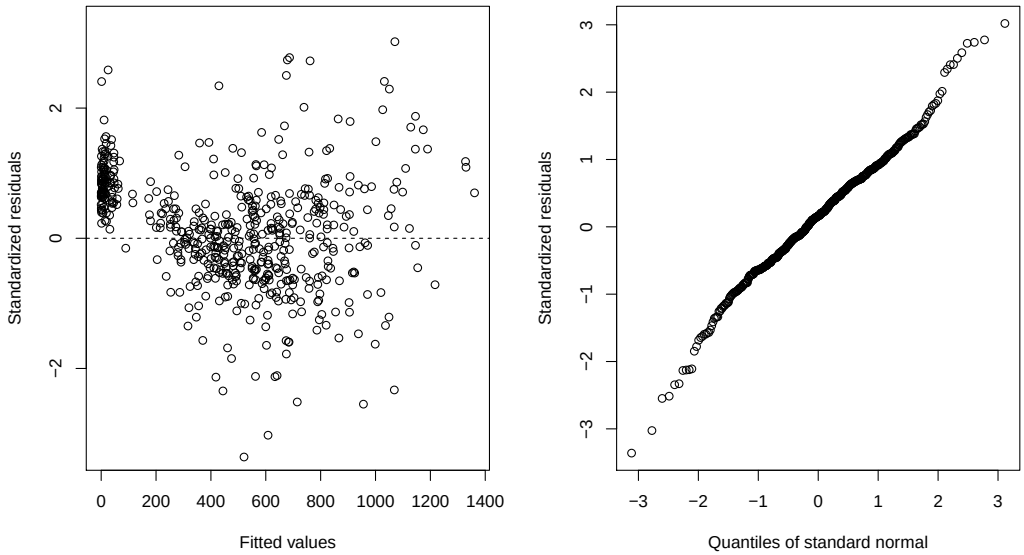


Figure 1: Standardized residuals versus fitted values and normal plot residuals for the final PK model.

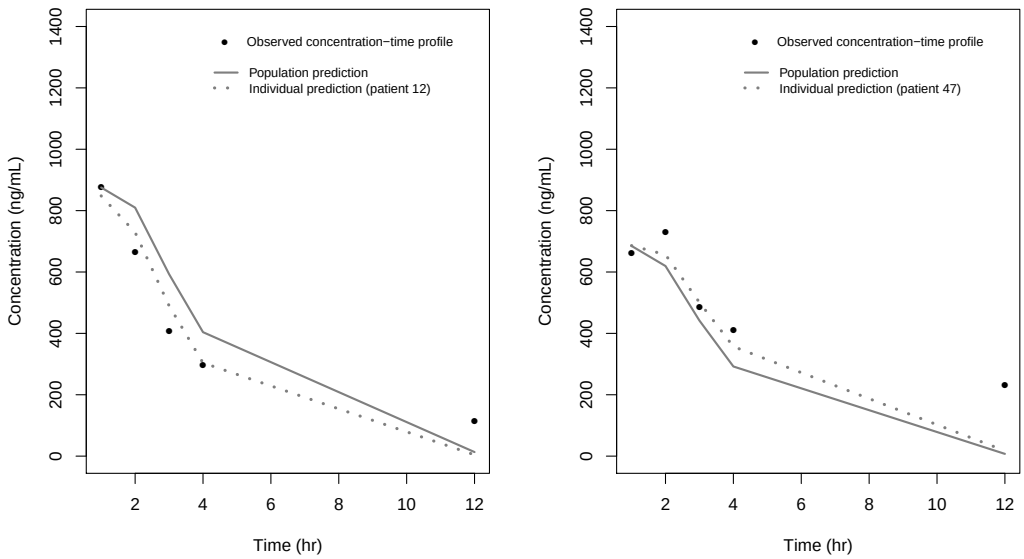


Figure 2: Observed and predict (population and individual) concentrations.

4. CONCLUSION

A PK model of CsA in the late stage of renal transplantation was successfully established. This model is simple and provides a useful tool that can be easily applied in clinical practice to estimate individual CsA dosing optimization in late stage renal transplant recipients by using conventional monitoring and to adjust dosing regimens with covariate factors (age, BSA and sex). Clinical model evaluation and further research in early stage of transplant are still required.

ACKNOWLEDGMENT

For the second author, the research was funded by FCT-Fundação para a Ciência e Tecnologia, Portugal, through the project UID/MAT/00006/2013.

References

[1] Bauer LA (2001.) *Applied clinical pharmacokinetics*. McGraw-Hill Professional.

[2] Pinheiro JC, Bates DM (2000). *Mixed-Effects Models in S and S-Plus*. Springer, New York.

[3] Davidian M, Giltinan DM (1995). *Nonlinear models for repeated measurement data*. Chapman and Hall, London.

[4] Rui JZ, Zhuo HT, Jiang GH, Chen G (1995). Evaluation of population pharmacokinetics of cyclosporin. A in renal transplantation patients with NONMEM. *Yao Xue Xue Bao* 30:241–7.

[5] Sherwin CMT, Kiang TKL, Spigarelli MG, Emsom MHM (2012). Fundamentals of population pharmacokinetic modeling: validation methods. *Clin Pharmacokinet* 51:573–590.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

JOINT MODELLING FOR LONGITUDINAL AND TIME-TO-EVENT IN HEALTH SCIENCES: WHERE WE ARE AND POSSIBLE EXTENSIONS

Inês Sousa[1,2]

[1]CBMA, University of Minho
[2]Department of Mathematics and Applications, University of Minho

ABSTRACT

Joint models in statistical context refer to model the joint distribution of more than one stochastic process with random variability. In particular, joint models of a response repeatedly measure on several subjects and time to an event of interest, is of main interest in health sciences. Although, in health sciences many data are produced with an association between these two processes, it is not that common to see in scientific publications the application of these models. The lack of software with the implementation of these models is one problem identified. These models are computationaly intensive to run, mainly due to the limitations when a more complex structure is impose. In this work a review on joint models for these data is developed and several extensions to the current implementations is explored.

Keywords and key sentences: longitudinal; survival; joint models.

1. INTRODUCTION

In health sciences, it is common to collect for several individuals repeated measures in time for responses of interest (longitudinal data) and, simultaneously, to be interested in the time until the occurrence of a clinical event (time-to-event data). Repeated measurements are analyzed with the so called longitudinal statistical models, that assume correlation between measurements of a same patient and independence between patients. The time-to-event response variable is analyzed within the survival models. However, when longitudinal responses are associated with the survival time it is known that independent longitudinal and survival analyses give biased results [1]. For example, women with breast cancer having a worst biomarker progression are more likely to die, or relapse, earlier. Therefore, if a longitudinal analysis is done only on the biomarker ignoring time-to-event, what we get to observe after deaths are the "best" patients, overestimating progression effect.

In the presence of association between the two processes, the observed longitudinal data is not a random sample of the population, they are realizations of a conditional distribution of the longitudinal variable, given they have survived up to a time point [2]. Therefore, longitudinal and time-to-event variables should be modeled within thei joint distribution. In the last

years, it has increased the interest on joint models for longitudinal and time-to-event data in health sciences [3,4,5,6,7].

In this work a review on joint models for longitudinal data and time-to-event is developed. The different approaches and their implementations are summarised. Although there is a very small number of software implementations for these models in the area of health sciences, it is not common to see publications using these models to analyse data with these characteristics. Joint modelling of longitudinal and survival data is already recognized by the medical community as a powerful tool. We will cover several publications that make use of these models in the area of health sciences. After, we will make some proposals for extensions of these models from the longitudinal and survival perspectives.

Although there has been a proliferation of joint models for longitudinal and survival data, all these have been based in the shared random effects model, highly computationally intensive.

2. Joint Modelling

In the 80's several authors discuss the importance of modelling longitudinal progression in the presence of dropout data, by modelling both processes [8]. Joint models were then motivated by the effect of a longitudinal variable on disease risk [9,10], assuming a common random effect in both longitudinal and survival models. Later a more complex structure for the random effects, including an unobserved stationary Gaussian process in a longitudinal linear predictor, was proposed [7]. Some overviews of these models have been given in [1] and [2].

Joint models have been extended to consider multiple longitudinal outcomes, modelling flexibly the subject specific longitudinal profiles, allowing for competing risks [11,12]. Although the models with shared random effects have been widely used, in practice these models are computationally demanding, due to high dimensional integrals that need numerical integration over random effects. Approximation methods have been proposed [5] but implementation is still difficult unless the random effects are of low dimension. Therefore, applications to fit joint models have been reduced mainly to random intercept and slope models. This applies to available software in R to fit these models (joineR and JM).

3. Extensions to Joint Modelling

A transformation Gaussian joint model was proposed [6], which is empirically more interpretable and straightforward to make inference on. In this model a multivariate structure is proposed for the longitudinal response and a log-Gaussian distribution for the time-to-event. Therefore, in this model the association is defined directly from the correlation of the multivariate distribution. This model does not need numerical approximations and is feasible with high dimensional variables.

Under any statistical model it is common to think of individual predictions for future measurements. In a longitudinal setting a dynamic process for individual predictions of longitudinal measurements has been developed [13]. We will also look to different ways of establishing predictions for longitudinal and survival times under a joint model.

Usually in survival data it is assumed right censoring data, when individuals leave the study for known or unknown reasons. In survival studies, it is common to observe left truncated data, where we do not have complete information on the number of individuals that experience

the event before a certain time point. We propose in this work to extend the transformation joint model to incorporate these survival extensions.

Going further on other types of censoring, we will develop the joint model incorporating interval censoring. For example, when the only information doctors have is that, in previous appointment patient was alive and the next appointment was dead. Meaning, failure occur during an interval of time. For the likelihood function this will contribute with two terms, a difference between two distribution functions.

Having in mind these possible extensions we will have the opportunity to extend the transformation Gaussian joint model to other survival settings as competing risks and cure models. Joint models allowing for competing risks have been studied before but always under the standard shared random effects joint model.

ACKNOWLEDGMENT

This work was supported by the strategic programme UID/BIA/04050/2013 (POCI-01-0145-FEDER-007569) funded by national funds through the FCT I.P.,by the Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) and by the ERDF through the COMPETE2020 - Programa Operacional Competitividade e Internacionalização (POCI).

References

[1] Tsiatis AA, Davidian M (2004), Joint modeling of longitudinal and time to event: An overview, *Statistica Sinica*, 14, 793–818

[2] Sousa I (2011), A review on joint modeling of longitudinal measurements and time-to-event, *REVSTAT-Statistical Journal*, 9 (1), 57–81

[3] Barret J, Diggle P, Henderson R, Taylor-Robinson D (2015), Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference, *Journal of the Royal Statistical Society-series B*, 77 (1), 131–148

[4] Proust-Lima C, Séne M., Taylor JMG, Jacqmin-Gadda H (2014), Joint latent class models for longitudinal and time-to-event data: a review, *Statistical Methods in Medical Research*, 23(1), 74–90

[5] Rizopoulos D (2012), Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule, *Computational Statistics and Data Analysis*, 56, 491–501

[6] Diggle PJ, Sousa I, Chetwynd AG (2008) Joint modelling of repeated measurements and time-to-event outcomes: The fourth Armitage lecture, *Statistics in Medicine*, 27 (16): 2981–2998

[7] Henderson R, Diggle P, Dobson A (2000), Joint modelling of longitudinal measurements and event time data, *Biostatistics* 1 (4) : 465–480

[8] Wu M, Carol R (1988) Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process, *Biometrics*, 44, 175–188

[9] Faucett CJ, Thomas DC (1996) Simultaneously modeling censored survival data and repeated measured covariates: A Gibbs sampling approach, *Statistics in Medicine*, 15, 1663–1685

[10] Wulfshon MS, Tsiatis AA (1997), A joint model for survival and longitudinal data measured with error, *Biometrics* 53 : 330–339

[11] Li N, Elashoff R, Li G (2009) Robust Joint Modeling of Longitudinal Measurements and Competing Risks Failure Time Data, *Biometrical Journal*, 51, 19–30

[12] Hu W, Li G, Li N (2009), A Bayesian approach to joint analysis of longitudinal measurements and competing risks failure time data, *Statistics in Medicine*, 28, 1601–1619

[13] Diggle PJ, Sousa I, Asar O (2015) Real-time monitoring of progression towards renal failure in primary care patients, *Biostatistics*, 16 (3), 522–536

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

ANALYSIS OF CLUSTERED ORDINAL SPATIAL PERIODONTAL DATA USING A NON-PARAMETRIC SPATIAL MODEL FOR INDEPENDENT LATTICES

Rui Martins[1], Vanda Inácio de Carvalho[2]

[1]Centro de Investigação Interdisciplinar Egas Moniz (CiiEM), Escola Superior de Saúde Egas Moniz e Instituto Universitário Egas Moniz, Monte de Caparica, 2829-511, Portugal: ruimartins@egasmoniz.edu.pt
[2]School of Mathematics, University of Edinburgh, EH9 3FD, Scotland, UK
vanda.inacio@ed.ac.uk

ABSTRACT

Periodontal disease (PD) is one of the major dental diseases that affect human populations worldwide at high prevalence rates. Chronic periodontitis (CP) is a prevalent inflammatory condition that results in the loss of tooth supporting connective tissue and alveolar bone. If untreated, is a major cause of tooth loss in adults.

Clinical attachment level (CAL), which measures the vertical distance between the cement-enamel junction and the lowest point (the base of the probable pocket) using a periodontal probe in an ordinal scale, is the most popular measure to assess periodontal disease's severity. Measurements obtained from a specific patient are intrinsically clustered within the mouth and proximal tooth sites have similar CAL values (spatial correlation) in comparison with sites that are further apart.

We consider here a non-parametric Bayesian approach for this ordinal measurements spatially clustered in separated lattices (i.e. mouths). Spatial association is incorporated considering the probit stick-breaking framework [**?**], which induces a natural clustering in the data allowing us to infer both teeth sites and groups of individuals with analogous PD status.

Keywords and key sentences: Periodontal disease; Stick-breaking prior; Spatial association; Independent lattices; Latent variable; Probit model.

1. INTRODUCTION

Chronic periodontitis is a prevalent condition and, if untreated, is a major cause of tooth loss in adults. Current techniques have limited applications, with somewhat unpredictable outcomes. Particularly, when the disease is widely distributed throughout the mouth, they are unable to fully reconstruct the periodontal tissues [**?**].
In the period 2009–2012, 8.9% of US adults had severe periodontitis. Overall, 46% of adults are disease's carriers in any one of its stages and 19.3% of sites (37.4% teeth) have CAL

$\geq 3mm$. [**?**]. The World Health Organization (WHO) recommends the integration of preventive strategies based on the most common risk factors in the public health practices and also recommends the establishment of a surveillance system for measuring the progress in the control of periodontal disease. According to a national cross-sectional study from the Portuguese Ministry of Health, the adult population estimated Periodontitis prevalence, at a confidence level of 95%, is $(32.4\%, 98.1\%)$ [**?**].

The progress of the periodontal disease results in irreversible destruction of the marginal alveolar bone, loss of associated periodontal ligament, and the apical migration of the junctional epithelium. Together these features are referred to as "loss of attachment". This, generally results in the formation of a pathologic periodontal probing depth, but may also cause gingival recession. CAL, which measures the distance between the cement-enamel junction and the lowest point using a periodontal probe, can be used as a criterion for the assessment of the severity of periodontal disease [**?**]. The probed tooth site level measures (6 per tooth) are rounded numbers (in millimetres) representing some ordering of the underlying PD progression but obtained manually. Because of that they are prone to errors of different kinds.

Modelling of variables observed at locations on a spatial lattice has been widely investigated and the family of conditionally autoregressive (CAR) models is a popular tool for analysing such data. However, there are many situations where the responses are observed in separated lattices for each subject, which is considerably different from the traditional setting (e.g. disease mapping) where several individuals are observed at each spatial location.

We consider here a non-parametric Bayesian approach for these ordinal measurements spatially clustered in separated lattices (i.e. mouths). Spatial association is incorporated considering the probit stick-breaking framework [**?**], which induces a natural clustering in the data allowing us to infer both teeth sites and groups of individuals with analogous PD status. Besides that we are interested in evaluate the covariates that can act as predictors of the CAL. The proposed methodology is illustrated using a real dataset.

2. Periodontal data

Our dataset comes from patient appointments at Clínica Universitária Egas Moniz. In this respect our data set differs from what is commonly seen, since it is not either (i) the result of a clinical trial (e.g. [**?**]) or (ii) the patients have been specifically recruit for that as in [**?**]. The $n = 51$ patients were at least 35 years old and all had chronic periodontitis. Each patient was examined a first time and returned two weeks later for a treatment. During a periodontal exam, CAL is usually measured at six sites for each of the 28 teeth (excluding the third molars, i.e., the wisdom teeth) corresponding to the mesio-buccal, mid-buccal, disto-buccal, mesio-lingual, mid-lingual and disto-lingual sites. An individual with no missing teeth has $m = 168$ ordinal measurements taking the values $k = 0, 1, \ldots, K-1$, with zero representing a PD–free site. Measures inside each mouth are highly correlated and clustered. The subject-level covariates included were age (years), gender (0=female, 1=male), smoking status (0=nonsmoker, 1=smoker), diabetes type II (0=no, 1=yes), arterial hypertension (0=no, 1=yes) and the total number of teeth per patient. Regarding the site-level covariates we considered whether the site is in a gap or not, i.e. if it is a mesio-buccal, disto-buccal, mesio-lingual or disto-lingual site (registed as 1), or if it is mid-buccal or mid-lingual (registed as 0). Finally, the jaw indicator (0=mandible; 1= maxilla) is also available.

We denote $y_i(s_j)$ the ordinal CAL measure at site s_j for the ith subject, $i = 1, \ldots, n$ and $j = 1, \ldots, m$; n and m are, respectively, the individuals and sites total. $y_i(\boldsymbol{s})$ represents the response vector for the ith subject with $\boldsymbol{s} = (s_1, s_2, \ldots, s_m)$. We will denote the $n \times p$ matrix of possible regressors by X, with the ith row, x_i, recording p covariates for the ith patient. Z will denote the $m \times q$ matrix of site level covariates, with the jth row, z_j, recording q covariates at site s_j.

3. Non-parametric spatial model

We want to capture the latent disease status so that we can obtain the supposed clustering of tooth sites and subjects with similar CAL level while taking into account the spatial correlations. Modelling ordinal data is naturally introduced if one considers an underlying continuous latent variable, $y_i^*(s_j)$. The ordinal outcome links to the latent variable through a cutpoints set. The probability are then represented by the probability that this latent variable belongs to a given interval defined by those cutpoints.
We set up an ordinal regression for $y_i(s_j)$ on covariates x_i and z_j. The CAL level probability $\Pr[y_i(s_j) = k]$ is represented as the probability that the continuous latent variable, $y_i^*(s_j)$, falls into the interval $[a_k, a_{k+1}[$. A patient-specific random effect site-dependent, $u_i(s_j)$, induces the correlations. Thus

$$y_i(s_j) = k \quad \text{if} \quad a_k \leq y_i^*(s_j) < a_{k+1}, \quad k = 0, 1, \ldots, K-1 \tag{1}$$

$$y_i^*(s_j) = x_i\boldsymbol{\beta} + z_j\boldsymbol{\gamma} + u_i(s_j) \tag{2}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the parameter vectors of dimension p and q, respectively. We assume that $u_i(s_j) \sim G$, being G a random probability measure with an appropriate prior distribution, $\mathcal{G}$. We will consider here the so-called probit-stick-breaking process (PSBP; [**?**]) for the unknown distribution G, i.e., G can be represented as

$$G = \sum_{h=1}^{\infty} \pi_h(s_j)\delta_{\theta_h}, \quad \pi_h(s_j) = \Phi(\alpha_{h_j}) \prod_{l<h} \left(1 - \Phi(\alpha_{l_j})\right), \quad \theta_h \overset{iid}{\sim} G_0, \quad j = 1, \ldots, m \tag{3}$$

where G_0 is the base measure. In this case we chose $G_0 \equiv N(0, 1)$. Here $\Phi(.)$ denotes the cumulative standard normal distribution. δ_θ is a degenerate distribution with all its mass at the atom θ. Notice that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ do not include an intercept parameter, because it is already implicitly in θ_h.
Here, the weights, $\pi_h(s_j)$, are generated from probit transformations of Gaussian-distributed α_{h_j}, which contrasts with the traditional construction that specifies a Beta$(1, \lambda)$ prior. Also, this formulation allows the weights, $\pi_h(s_j)$, to be dependent on a set of covariates by letting the parameter α_{h_j} being dependent on the covariates measured for y_{ij}. The summation representing G can be finite (known, or unknown) or infinite and in [**?**] is shown that $\sum_{h=1}^{\infty} \pi_h(s_j) = 1$, almost surely . The spatial dependence is incorporated considering for each mixture component, h, that $\alpha_h = (\alpha_{h_1}, \alpha_{h_2}, \ldots, \alpha_{h_m})^\top \sim \mathcal{N}_m(\mathbf{0}, \Sigma)$, where Σ is the corresponding $m \times m$ covariance matrix.
The choice as prior distribution of the PSBP has an appealing propriety. Marginalizing out the random probability measure, G, one induces dependencies between the random-effects variables and implies that the distribution of the latent $y_i^*(s_j)$ is an infinite mixture of Gaussian distributions with location parameters $(x_i\boldsymbol{\beta} + z_j\boldsymbol{\gamma} + \theta_h)$ and scale σ^2. If one considers a latent indicator, $v_i(s_j)$, such that $v_i(s_j) = h$ if the response for site s_j of patient i comes from the hth mixture component with probability $\pi_h(s_j)$, the conditional distribution of the latent $y_i^*(s_j)$ is a Gaussian, $(y_i^*(s_j)|\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta_h, v_i(s_j) = h) \sim \mathcal{N}(x_i\boldsymbol{\beta} + z_j\boldsymbol{\gamma} + \theta_h, \sigma^2)$. Thus, marginalizing with respect to $y_i^*(s_j)$, we have

$$\Pr\left(y_i(s_j) = k|\boldsymbol{\beta}, \boldsymbol{\gamma}, \theta_h, v_i(s_j) = h\right) = \Phi\left(\frac{a_{k+1} - x_i\boldsymbol{\beta} - z_j\boldsymbol{\gamma} - \theta_h}{\sigma}\right) - \Phi\left(\frac{a_k - x_i\boldsymbol{\beta} - z_j\boldsymbol{\gamma} - \theta_h}{\sigma}\right).$$

For reasons of identifiability σ is usually fixed to 1 and, without loss of generality, the cutpoints a_k can also be fixed as $a_0 = -\infty$, $a_K = \infty$ and $\{a_1, \ldots, a_{K-1}\} = \{-4, 0, 4, 8\}$ ([**?**] and [**?**]). The choice provides a wide support $]-\infty, -4]$ and $[8, +\infty[$ to the extreme events, i.e. "normal" and "missing", respectively, because most of the data are expected to have some proportion of both healthy and missing sites [**?**].

4. CONCLUSIONS

This work stems from the needs of the dental practice at Clínica Universitária Egas Moniz, which provided the context and motivation to better comprehend one of the world's most prevalent diseases. We presented here a framework to capture the state of the latent PD while allowing for the understanding of how tooth sites and subjects with similar disease status are clustered. Similar studies (e.g. [**?**]) had already point the advantages of considering flexible non-parametric approaches to treat this kind of data.
Preliminary results of our models showed that the variables age, gender, diabetes type II and hypertension are not significant.

ACKNOWLEDGMENT

We acknowledge Clínica Universitária Egas Moniz for the dataset. This work was partially funded by Fundação para a Ciência e a Tecnologia (FCT) project UID/BIM/04585/2016.

References

[1] Bandyopadhyay, D. and Canale, A. (2016). Non-parametric spatial models for clustered ordered and periodontal data. *Journal of the Royal Statistical Society, Ser. C*, 65, 619–640.

[2] Calado, R., Ferreira, C., Nogueira, P. and Melo, P. (2015). III Estudo Nacional de Prevalência das Doenças Orais. *Direcção Geral da Saúde*.

[3] Eke, P., Dye, B. and Wei, L. *et al.* (2015). Update on prevalence of periodontitis in adults in the United States: NHANES 2009 to 2012. *Journal of periodontology*, 86(5), 611–622.

[4] Fernandes, J., Wiegand, R., Salinas, C., *et al.* (2009). Periodontal disease status in Gullah African Americans with type 2 diabetes living in South Carolina. *Journal of periodontology*, 80(7), 1062–1068.

[5] Hughes, J. (2015). Periodontium and Periodontal Disease. In Vishwakarma, A., Sharpe, P., Shi, S. and Ramalingam, M. (eds) *Stem Cell Biology and Tissue Engineering in Dental Sciences*, Chp 34, 433–444, Academic Press.

[6] Kottas, A., Müeller, P. and Quintana, F. (2005) Nonparametric Bayesian modeling for multivariate ordinal data. *J. Comput. Graph. Statist.*, 14(3), 610–625.

[7] Leon-Novelo, L., Zhou, X., Bekele, B. and Müller, P. (2010). Assessing toxicities in a clinical trial: Bayesian inference for ordinal data nested within categories. *Biometrics*, 66(3), 966–974.

[8] Mdala, I., Haffajee, A., Socransky, S. *et al.* (2012). Multilevel analysis of clinical parameters in chronic periodontitis after root planing/scaling, surgery, and systemic and local antibiotics: 2-year results. *Journal of oral microbiology*, 4(1), 17535.

[9] Rodríguez, A. and Dunson, D. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6(1), 145–177.

[10] Smiley, C., Tracy, S. and Abt, E. *et al.* (2015). Evidence-based clinical practice guideline on the nonsurgical treatment of chronic periodontitis by means of scaling and root planing with or without adjuncts. *The Journal of the American Dental Association* 146(7), 525–535.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

PERCEÇÃO PARENTAL DO PESO E ESTILOS DE VIDA DOS ADOLESCENTES - UMA APLICAÇÃO DE MEDIDAS DE CONCORDÂNCIA ENTRE INQUÉRITOS

Elsa Silva[1], Augusta Gama[2] e Marília Antunes[3]

[1] Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[2] Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[3] Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

RESUMO

Neste trabalho, são analisados 664 pares de inquéritos aplicados a adolescentes e aos seus progenitores, visando aspetos relativos a peso, saúde e estilos de vida. O objetivo é quantificar o grau de concordância das respostas dadas por ambos, enquanto medida indicadora da perceção parental do peso e estilos de vida dos adolescentes.

Palavras chave: Medidas de concordância.

1. INTRODUÇÃO

Segundo a Organização Mundial de Saúde, a prevalência do excesso de peso e de obesidade em crianças e adolescentes, de 5 a 19 anos, aumentou de 4% em 1975 para 18% em 2016, existindo, nesse ano, mais de 340 milhões de crianças e jovens com excesso de peso ou obesidade. Entre os países mais afetados encontra-se Portugal em que, de acordo com um estudo de Sardinna et al. [1], 17,0% das raparigas e 17,7% dos rapazes, de 10 a 18 anos, sofriam de excesso de peso enquanto 4,6% das raparigas e 5,8% dos rapazes, do mesmo grupo etário, sofriam de obesidade.

O peso em excesso tem efeitos prejudiciais para a saúde e bem-estar, a curto e longo prazo, sendo um fator de risco para diversas doenças crónicas como a diabetes, doenças cardiovasculares, hipertensão e cancro, sendo os baixos níveis de atividade física, os elevados níveis de sedentarismo tecnológico e a alimentação desadequada apontados como alguns dos hábitos que os jovens têm hoje em dia e que mais contribuem para o aumento da prevalência do excesso de peso e de obesidade na adolescência.

Tendo os pais um papel de grande influência nos hábitos que os filhos adquirem ao longo da vida, sobretudo durante a infância e adolescência, torna-se clara a importância de uma correta perceção dos pais em relação aos hábitos dos filhos, principalmente quando estes contribuem para um aumento de peso prejudicial para a saúde.

Neste trabalho, analisamos 664 pares de inquéritos aplicados a jovens, com idades entre 12 e 18 anos, e a um dos seus progenitores (ou, caso o jovem não habite com estes, o adulto responsável), com o objetivo de quantificar o grau de concordância das respostas dadas por ambos.
Os inquéritos em causa foram elaborados no âmbito do Programa Nacional de Saúde Escolar, com o objetivo de caracterizar o ambiente familiar de jovens pertencentes a escolas com elevada prevalência de casos de excesso de peso e obesidade.
Os questionários aplicados aos jovens abordam questões relativas a atividade física, saúde, hábitos alimentares e de sono, na sua maioria em forma fechada (resposta qualitativa ou ordinal). Algumas questões são relativas à perceção que os jovens têm de si mesmos, sobretudo no que respeita à sua saúde, peso e apetite. Os questionários aplicados aos progenitores são diferentes, tanto no número de questões colocadas como na forma em que são apresentadas, ainda que dirigidas aos mesmos aspetos. A maioria das questões são dirigidas para a preocupação e atenção que têm com a saúde, alimentação e hábitos dos filhos.
Ambos os questionários não foram validados para a população portuguesa, no entanto, foram construídos com questões retiradas de questionários internacionais e nacionais validados como: Youth Risk Behavior Survey (YBRS), IPAQ, Aventura Social/KIDSCREEN e Estudo Nacional de Prevalência de Obesidade Infantil em Portugal, alterações de 2002 a 2009.

2. MÉTODOS

Com base nas respostas dadas pelos pais e filhos a questões relativas ao peso, saúde e estilos de vida dos adolescentes, como hábitos sedentários, alimentares e de atividade física, é possível determinar o grau de concordância entre essas respostas construindo, para o efeito, medidas de concordância baseadas numa extensão do Coeficiente de Gower proposta por Podani [2].
Para que estas medidas sejam aplicadas de forma correta, é necessário identificar quais as questões que são semelhantes nos inquéritos dos adolescentes e dos pais e a escala em que as respostas se encontram. No caso dos pares de questões identificadas como semelhantes não possuírem o mesmo tipo de respostas possíveis (seja em escala ou em número), é necessário que se comece por hamonizar as respostas possíveis de ambas as questões, agrupando-as da mesma forma nos dois questionários.
Concluída a fase de harmonização dos questionários, dependendo da natureza da escala de cada questão, é escolhida a medida de concordância que se considera adequada. Assim, para cada par de respostas $(pai_j, filho_j)$:

- se questão i for de resposta/escolha múltipla (nominal) com K níveis e a possibilidade do inquirido registar mais de uma escolha, a medida de concordância é:

$$s_i(p_j, f_j) = \frac{\sum_{k=1}^{K} s_{ik}(p_j, f_j)}{K},$$

em que, para cada nível k:

$$s_{ik}(p_j, f_j) = \begin{cases} 1 & \text{se} \quad p_{ikj} = f_{ikj}, \\ 0 & \text{se} \quad p_{ikj} \neq f_{ikj}. \end{cases},$$

onde p_{ikj} representa a resposta ao nível k da questão i registada pelo pai, definindo-se f_{ikj} de forma análoga.

- se questão i for quantitativa:

$$s_i(p_j, f_j) = 1 - \frac{\mid p_{ij} - f_{ij} \mid}{R_i},$$

em que:

$R_i = \max\big(\max_j(p_{ij}), \max_j(f_{ij})\big) - \min\big(\min_j(p_{ij}), \min_j(f_{ij})\big)$.

- se questão i for ordinal:

$$s_i(p_j, f_j) = 1, \quad se \;\; r_i(p_j) = r_i(f_j)$$

$$s_i(p_j, f_j) = 1 - \frac{\mid r_i(p_j) - r_i(f_j) \mid - \big(T_i(p_j) - 1\big)/2 - \big(T_i(f_j) - 1\big)/2}{\max\{r_i\} - \min\{r_i\} - \big(T_{i,\max} - 1\big)/2 - \big(T_{i,\min} - 1\big)/2}, \;\; se \;\; r_i(p_j) \neq r_i(f_j).$$

em que:

  - $r_i(p_j)$: *rank score* da resposta do pai_j;
  - $r_i(f_j)$: *rank score* da resposta do $filho_j$;
  - $T_i(p_j)$: nº de pais e filhos que têm igual *rank score* que o pai_j;
  - $T_i(f_j)$: nº de pais e filhos que têm igual *rank score* que o $filho_j$;
  - $T_{i,\max}$: nº de pais e filhos que têm *rank* máximo= $\max\{r_i\}$;
  - $T_{i,\min}$: nº de pais e filhos que têm *rank* mínimo= $\min\{r_i\}$.

Por fim, é aplicado o Coeficiente de Gower, que vai agregar os níveis de concordância obtidos e que permitirá quantificar o grau de concordância global entre as respostas dos filhos e dos pais. Os valores destas medidas variam entre 0 e 1, apontando, respetivamente, para ausência de concordância e concordância total. Assim sendo, valores mais elevados desta medida revelam uma boa perceção parental do peso e estilos de vida dos adolescentes.
Como trabalho futuro serão construídos modelos de regressão beta [3] para conhecer o papel das características dos pais e do meio familiar na concordância entre pais e filhos.

AGRADECIMENTOS

O trabalho de Marília Antunes é financiado por Fundos Nacionais através da FCT no âmbito do projeto UID/MAT/00006/2013.

Referências

[1] Sardinna, L., Santos, R., Vale, S., Silva, A., Ferreira, J., Raimundo, A., Moreira, H., Baptista, F. , Mota., J. (2011). Prevalence of overweight and obesity among Portuguese youth: A study in a representative sample of 10–18-year-old children and adolescents. *International Journal of Pediatric Obesity*, 6, (2-2): e124-8.

[2] Podani J. (1999). Extending Gower's General Coefficient of Similarity to Ordinal Characters. *Taxon*, Vol. 48, No. 2, 331-340.

[3] Ferrari, S. , Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, Vol. 31, No. 7, 799–815.

AVALIAÇÃO DA ATIVIDADE DO LÚPUS: SLEDAI VS EVA

Ana Cristina Matos [1], Carla Henriques [2] e Diogo Jesus [3]

[1] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu; Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS), amatos@estv.ipv.pt

[2] Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu, Centro de Matemática da Universidade de Coimbra (CMUC), Centro de Estudos em Educação, Tecnologias e Saúde (CI&DETS), carlahenriq@estv.ipv.pt

[3] Clínica de Lúpus, Serviço de Reumatologia, Centro Hospitalar e Universitário de Coimbra, jesus.p.diogo@gmail.com

RESUMO

Lúpus é uma doença crónica autoimune com atividade clínica variável, podendo apresentar atividade persistente, ou evoluir com períodos de agudização intercalados com períodos de inatividade. A avaliação da atividade de doença é deveras crucial na determinação da terapêutica a adotar bem como monitorização da sua eficácia. Este trabalho foca-se na comparação de duas medidas de avaliação da atividade desta doença: a avaliação global da atividade pelo médico numa escala visual analógica – EVA e a avaliação produzida com base num índice obtido por soma de valores que pontuam a presença de manifestações clínicas e serológicas – o SLEDAI. O estudo envolve registos destas avaliações de pacientes que foram seguidos durante dois anos. É analisada a correlação e a concordância entre os resultados dos dois métodos. Adicionalmente é avaliada a capacidade do SLEDAI em detetar melhorias/agravamentos aferidos pela avaliação médica.

Palavras e frases chave: Lúpus Eritematoso Sistémico, concordância, curvas ROC.

1. INTRODUÇÃO

O Lúpus Eritematoso Sistêmico (LES ou apenas lúpus) é uma doença crónica, multissistémica, com evolução clínica variável podendo apresentar atividade persistente ou evoluir por períodos de surtos alternados com remissões. Dada a sua complexidade, têm sido desenvolvidos vários índices para quantificar a atividade de doença, valorizando de forma distinta diferentes manifestações clínicas. O SLEDAI – *Systemic Lupus Erythematosus Disease Activity Index*, é dos índices mais utilizado na prática clínica, com uma pontuação que pode variar entre 0 e 105, resultando da soma de itens com diferentes ponderações, consoante a presença/ausência de determinadas manifestações clínicas e laboratoriais.

Como avaliação de referência (*gold standard*) é utilizada a EVA, uma Escala Visual Analógica em que o médico quantifica a atividade global de doença entre 0 e 30 pontos.
O nosso estudo envolve 279 pacientes que foram seguidos entre janeiro de 2014 a dezembro de 2016.
A análise longitudinal da atividade da doença de cada paciente foi sumariada recorrendo ao cálculo da média ajustada do período em estudo uma vez que o intervalo entre consultas é variável. O cálculo desta estatística sumária para os valores atribuídos ao EVA e para a pontuação do SLEDAI em cada paciente, permitiu calcular a associação entre os dois scores de avaliação. Também foi avaliada a correlação entre os dois scores para cada consulta.
A evolução clínica dos doentes foi feita comparando estes scores entre duas consultas consecutivas (melhoria, agravamento ou mesmo estado). Estudou-se a associação e a concordância entre os scores em estudo quanto ao resultado desta classificação, utilizando o teste do Qui-quadrado, o coeficiente Kappa de Cohen e coeficiente Kappa de Cohen ponderado.
Devido à natureza rígida do cálculo do SLEDAI, é por vezes posta em causa a capacidade deste índice em detetar melhorias/agravamentos identificadas pelo médico - EVA. Neste trabalho é feito um estudo de modo a avaliar a capacidade do SLEDAI em diferenciar os doentes que melhoram/pioram de acordo com a EVA. Adicionalmente, estuda-se também a habilidade do SLEDAI para detetar melhoras/agravamentos clinicamente relevantes, sendo estes caracterizadas por uma variação no valor da EVA de pelo menos 10% da escala. Este estudo envolveu a construção de curvas ROC (Receiver Operating Characteristic) e sua análise, de forma a aferir a capacidade do SLEDAI para diferenciar doentes identificados com melhoras/agravamentos na EVA, relevantes ou não, assim como obter uma indicação do valor de variação no SLEDAI que melhor identifica a evolução na EVA.

2. RESULTADOS

Foram considerados registos de 279 doentes (86,1% do género feminino) seguidos por um período de dois anos na Clínica de Lúpus, Serviço de Reumatologia, Centro Hospitalar e Universitário de Coimbra.
A média ajustada dos índices EVA e SLEDAI apresentou uma correlação de Spearman de 0,824 ($p < 0{,}0005$). A correlação dos dois índices obtidos em cada visita, medida através do coeficiente de correlação de Spearman, apresenta valor mediano de 0.8407, com intervalo inter-quartil de 0.792-0.873.
Traduzindo a evolução do doente, da segunda para a primeira consulta, num agravamento (o doente piorou na segunda consulta, isto é, o score de atividade da doença aumentou), melhoria (o score de atividade da doença diminuiu) ou sem alteração (o score manteve-se), mediu-se o grau de associação entre as classificações obtidas a partir da EVA e do SLEDAI. Esta associação revelou-se estatisticamente significativa (teste Qui-quadrado, $p<0.0005$), no entanto a concordância, avaliada através do coeficiente Kappa de Cohen ponderado, apenas se pode considerar moderada [2] – Figura 1.

		Sledai			
		piorou	manteve	melhorou	Total
EVA	piorou	27	18	6	**51**
	manteve	9	108	9	**126**
	melhorou	3	37	59	**99**
Total		**39**	**163**	**74**	**276**

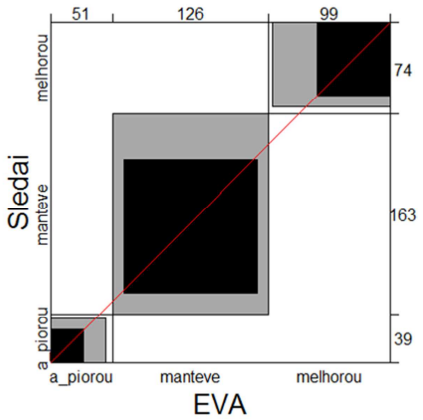


Figura 1: *Agreementplot* da evolução da primeira para a segunda consulta

Quando avaliadas as concordâncias entre duas consultas consecutivas, obtivemos um conjunto de valores para o coeficiente Kappa de Cohen e para o coeficiente Kappa de Cohen ponderado com intervalos interquartis de (0,47-0,69) e (0,51-0,69), respetivamente.
A avaliação da capacidade do SLEDAI para diferenciar os doentes que melhoram/pioram de acordo com a EVA, da primeira para a segunda consulta, foi feita através de curvas ROC (Figura 2). Considerou-se, pois, a diferença entre os scores SLEDAI das duas primeiras consultas e pesquisou-se qual a capacidade desta diferença para diferenciar doentes que, na perspetiva do clínico, isto é, de acordo com a EVA, teriam melhorado (Figura 2A), piorado, (Figura 2B), melhorado de forma relevante (Figura 2C) ou piorado de forma relevante (Figura 2D).

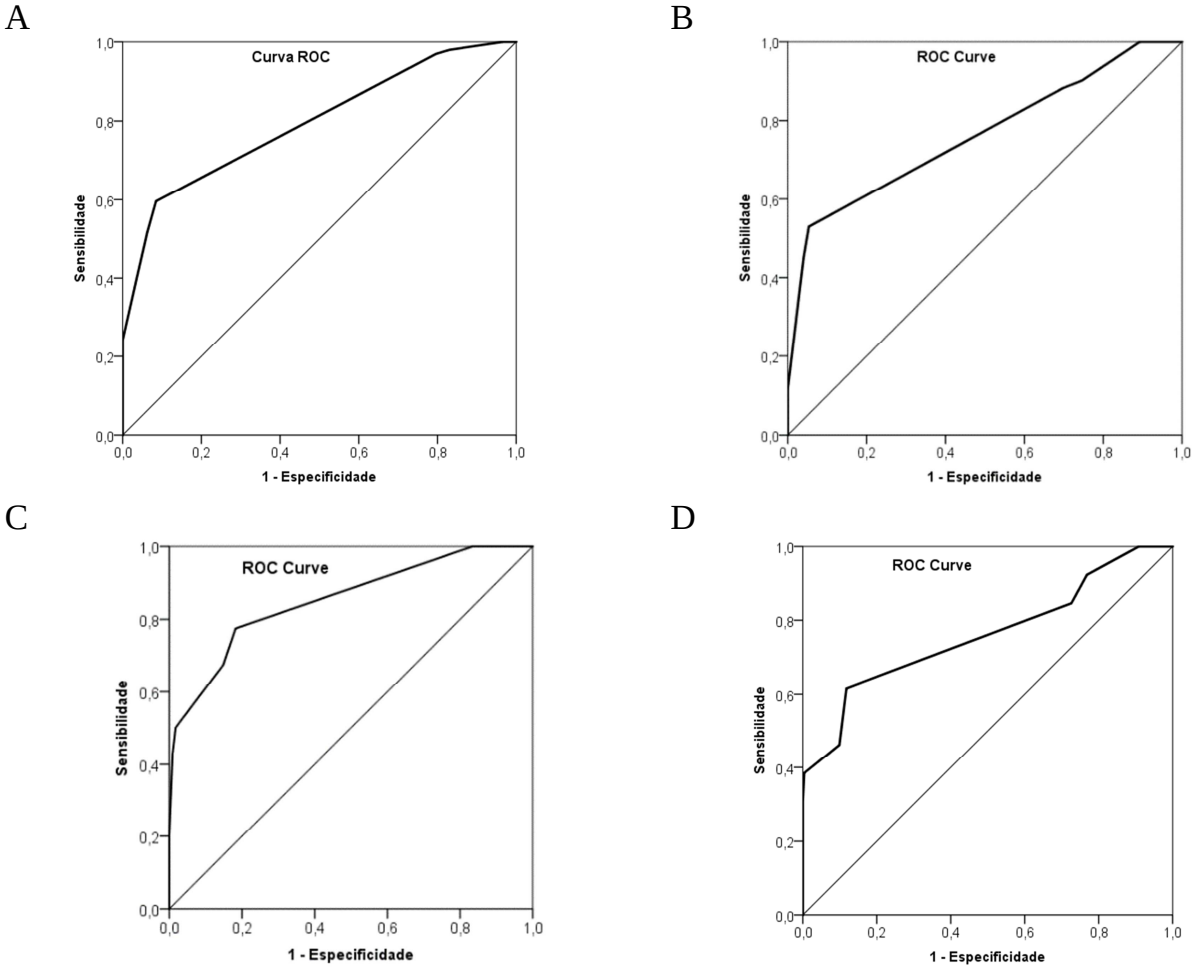


Figura 2: Capacidade da diferença nos scores SLEDAI para diferenciar doentes que, da primeira para a segunda consulta, de acordo com o EVA, melhoraram (A), pioraram (B), melhoraram de forma relevante (C) ou pioraram de forma relevante (D)

A curva ROC para a melhoria (Figura 2A) apresentou uma AUC de 0,795 (IC95%: 0,737 – 0,852), indicativa de que o SLEDAI tem uma capacidade aceitável [1] para distinguir os doentes que melhoraram na perspetiva do clínico. Claramente, o *cutoff* que representa o melhor compromisso entre sensibilidade e especificidade é dado por uma diferença no SLEDAI de um ponto, para o qual a sensibilidade é de apenas 0,596 e a especificidade de 0,915. Isto é, apenas 59,6% dos doentes que melhoraram na segunda consulta são identificados como tal pelo SLEDAI. Para a situação de agravamento (Figura 2B), como seria de esperar, a avaliação do SLEDAI, como índice diferenciador dos doentes que pioram, é idêntica. A curva ROC tem AUC=0,763 (IC95%: 0,68 – 0,845), o que mais uma vez se pode traduzir numa capacidade discriminativa aceitável. São também idênticos os valores de sensibilidade e especificidade obtidos para o melhor *cutoff* (correspondente a uma diferença de um ponto no SLEDAI),

respetivamente 0,529 e 0,947. Considerando uma melhoria ou um agravamento relevantes (correspondentes a uma variação de pelo menos 3 pontos no valor do EVA), as curvas ROC apresentam valores de AUC iguais a 0,853 (IC95%: 0,782 – 0,924) e 0,760 (IC95%: 0,596 – 0,924), respetivamente (Figuras 2C e 2D). Mais uma vez o ponto de corte sugerido pelas curvas ROC situa-se numa diferença no SLEDAI igual a um ponto, obtendo-se uma sensibilidade de 0,775 e uma especificidade de 0,818, no caso de uma melhoria relevante, e de 0,615 e 0,882, no caso de um agravamento relevante.

Em resumo, os valores obtidos para a sensibilidade indicam que o SLEDAI consegue identificar pouco mais de metade dos doentes que apresentaram efetivamente alguma melhoria. Naturalmente, tratando-se de identificar doentes com uma melhoria relevante, a performance do SLEDAI melhora, isto é, o SLEDAI diminui pelo menos um ponto em 78% dos doentes com melhorias relevantes. Para o agravamento, uma diferença de um ponto no SLEDAI consegue identificar 62% dos doentes com agravamento relevante.

3. CONCLUSÃO

Conclui-se que apesar de haver uma correlação forte entre as duas escalas de avaliação (rho de Spearman=0.83, $p<0.0005$) a concordância não se pode considerar elevada. Tendo a EVA como referência, e considerando no SLEDAI uma diferença de pelo menos 1 ponto, o SLEDAI só deteta cerca de metade dos casos de melhoras e também cerca de metade dos casos de agravamentos. O SLEDAI apresenta um desempenho limitado na deteção de uma mudança clinicamente relevante, o que sugere a necessidade de otimizar este índice de atividade de doença.

AGRADECIMENTOS

Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto UID/Multi/04016/2016. Agradecemos adicionalmente ao Instituto Politécnico de Viseu e ao CI&DETS pelo apoio prestado. Adicionalmente, este trabalho foi parcialmente apoiado pelo Centro de Matemática da Universidade de Coimbra - UID / MAT / 00324/2013, financiado pelo Governo Português através da FCT / MEC e co-financiado pelo Fundo Europeu de Desenvolvimento Regional através do Acordo de Parceria PT2020.

Referências

[1] Hosmer, D. W., Lemeshow, S. (2000). *Applied Logistic Regression* (2nd edition). New York: John Wiley & Sons.

[2] Landis JR, Koch GG. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159-174.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

HYPERSPECTRAL IMAGE CLASSIFICATION USING FUNCTIONAL DATA ANALYSIS

M. Oviedo de la Fuente[1,2], M. Febrero-Bande[2,1]

[1]Technological Institute for Industrial Mathematics, Spain
[2]MODESTYA group member, Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, Spain

ABSTRACT

In the first part of the work, we will review different models classification algorithms for the prediction of the future class of pixel in a hyperspectral image that have in common that make use of Functional Data Analysis (FDA). The advantage of FDA over classical model is that it is able to exploit this continuous nature of the information of spectral curves in a better way. The second part of the communication is devoted to the problem of variable selection. This work proposes a selection method that is designed to mixed covariates of different nature: scalar, multivariate, functional, etc. The proposal begins with a simple null model and sequentially selects a new variable to be incorporated into the final prediction model. The algorithm have showed quite promising results when applied to hyperspectral image classification.

1. INTRODUCTION

During the last years, the use of hyperspectral sensors has been extended to a great variety of applications such as discrimination among different land cover classes in remote sensing images, see Figure 1.

In multivariate data analysis, classical Machine Learning methods (random forest, neural networks, decision tree, SVM, see [1]) are used in hyperspectral image classification problems. However, usually these procedures are computed repeatedly to each of the spectral bands and they no consider the high correlation between consecutive spectra.

2. Functional Data Classification

Functional data analysis (FDA) is a branch of statistics that analyzes data providing information about curves, surfaces or anything else varying over a continuum. The continuum is often time, but may also be spatial location, wavelength, probability, etc.

- **Definition 2.1**. A random variable $\mathcal{X}$ is called a functional variable if it takes values in a functional space $\mathcal{E}$ –complete normed space–.
- **Definition 2.2**. A functional dataset $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ is the observation of n functional variables $\mathcal{X}_1, \ldots, \mathcal{X}_n$ identically distributed as $\mathcal{X}$.

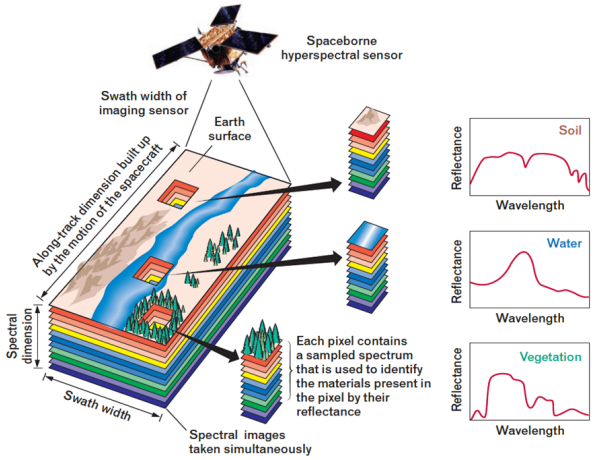


Figure 1: Example illustrating the problem in remotely sensed hyperspectral data analysis.

Bayes rule: Given a sample $\mathcal{X}$ of a functional variable, the aim is to estimate the posterior probability of belonging to each group:

$$p_g(\mathcal{X}) = P(Y = g|\chi = \mathcal{X}) = E(1_{Y=g}|\chi = \mathcal{X})$$

The classification rule is to assign a new functional observation that group with the maximum a posteriori probability: $\hat{Y} = arg\max \hat{p}_g(\mathcal{X})$. The estimate of the posterior probability $p_g(\mathcal{X})$ can be calculated using logistic regression or non parametric regression. In FDA, the large number of spectral bands (highly correlated) can be treated such as a continuous function, so that for each pixel has associated a function (spectral curve) with order in the bands. The objective is predict the associated class for each pixel considering the shape of each spectral curve.

We used a multiclass one-versus-one (majority voting) and one-versus-rest (maximum probability) functional GLM and GAM models in hyperspectral image classification. Techniques for FD representation, fixed basis (Fourier, Splines) or data driven basis (FPCA, PLS), are applied in order to reduce the dimension data. On the other hand, functional non-parametric classification by mean of k-nearest neighbors classifiers can help to discriminate the pixel class through the shape of the spectral curve.

3. Impact Points and Variable Selection

The variable selection problem in a general regression/classification model tries to find the subset of covariates that best predicts or explains a response. In the classical approach, the covariates and the response are scalar (or binary) and the model established among them is linear. In a general framework, this problem is even more important because now, in the Big Data era, huge amounts of information are available but the information is contained on variables of different nature: functional, scalar, directional, categorical, etc. Our purpose in the rest of this work is to provide an automatic procedure for selecting classification models with a subset of the available covariates that are of different nature (spatial -positions of pixels-, functional -spectral curves-, multivariate -sensitive impact points- or further covariates). Our aim is to select significant covariates for a general additive regression model (GAM) with categorical response Y:

$$Y_i = g^{-1}\left(\sum_{j=1}^{J} f_j\left(\mathcal{X}_i^{(j)}\right)\right) + \varepsilon_i, \quad i = 1, \ldots, n$$

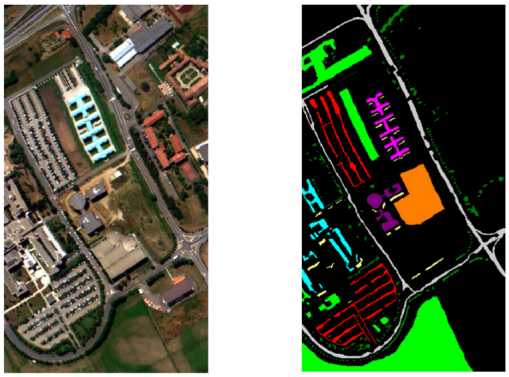
where g^{-1} is the inverse of the link function and the covariates are chosen from the set $S = \left\{\mathcal{X}^1, \mathcal{X}^2, \cdots, \mathcal{X}^k, \cdots\right\}$ of different potential covariates (functional, vectorial, ...). The notation $\mathcal{X}^{(j)}$ refers to the j-th covariate selected for the model. The number of variates can be extraordinarily large, so we intent to construct the regression model sequentially,

i.e. from the trivial model up to the one that includes all the useful information provided by the covariates in the set S.

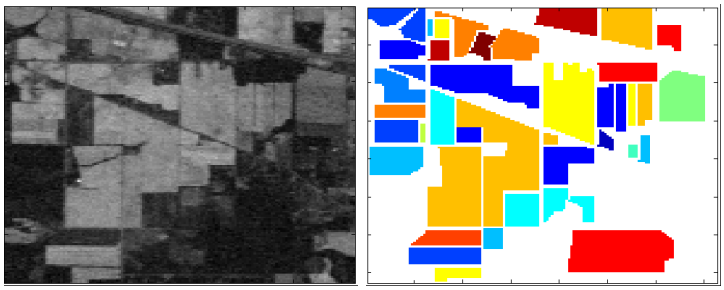
3. Results

The efficiency of the proposed methods is evaluated on bellow hyperspectral data sets and compare to some state-of-the-art hyperspectral image classification methods. The experiments are conducted on the the University of Pavia and Indian Pine images, respectively.

- **Pavia University**: This is a remote sensing image obtained by the 103—band ROSIS sensor from the University of Pavia, With a spatial dimension of 610 × 340 pixels.



- **Indian Pines**: This is a remote sensing image obtained by the AVIRIS 220-band and 145 × 145-pixel sensor taken over Northwest Indiana.



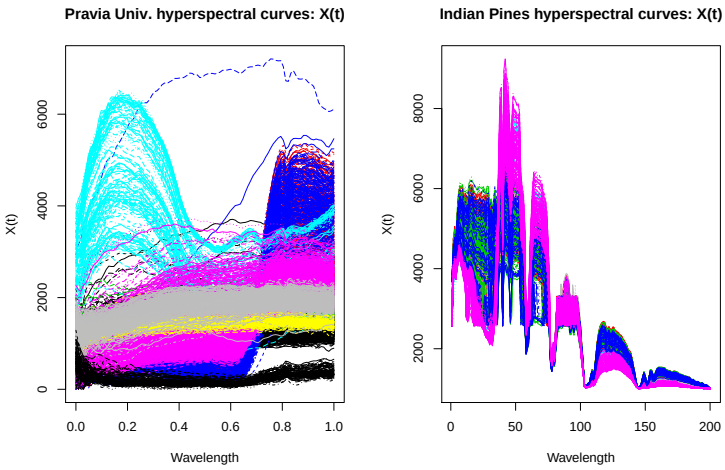


Figure 2: Hyperspectal curves of Univ. of Pavia and Indian Pines.

The algotithms compared are:

- FkNN: Functional k-nearest neighbors, see [2].

- RPART: Recursive partitioning and regression trees.
- Functional (linear or additive) logistic regression is apply (FGLM and FGAM models) using maximum probability (MaxProb) and majority voting (MajVot) schemes.
- Local maxima distance correlation (LMDC), see [3]. The impact points are selected using LMDC, then an additive logistic regression is applied.
- A modified algorithm proposed by [4] is applied.

In the University of Pavia image, 300 pixels for each class were chosen for training (see a sample of the curves in left panel of Figure 2), and the rest were used as a test set. For Indian Pines image, we select randomly 100 pixels for class for train the classifier (see right panel of Figure 2). The results are tabulated in Table 1.

Dataset	Univ. Pavia		Indian Pines	
Method	$X(t)$	$X(t)$, x-pos., y-pos.	$X(t)$	$X(t)$, x-pos., y-pos.
RPART	0.584	0.892	0.532	0.934
FkNN	0.708	NA	0.660	NA
FGLM+MaxProb	0.603	0.809	0.601	0.745
FGLM+MajVot	0.688	0.914	0.647	0.820
FGAM+MaxProb	0.709	0.988	0.731	0.928
FGAM+MajVot	0.748	0.985	0.747	0.926
LMDC+GAM	0.759	NA	0.717	NA
VS+FGAM	0.717	0.976	0.700	0.893

Table 1: Accuracy average over 25 runs. NA: Method not available.

4. Conclusion

Classification using majority voting improves the accuracy compared with maximum probability scheme. Techniques for variable selection are as good alternative to FGLM and FGAM procedures, so provide similar results and inform about the most predictive wavelength points of the spectral curve.

Keywords and key sentences: Hyperspectral imaging, functional classification, Kernel, principal component analysis, support vector machines.

ACKNOWLEDGMENT

The authors acknowledge financial support from Ministerio de Economía y Competitividad grant MTM2016-76969-P and European Regional Development Fund (ERDF).

References

[1] Melgani, F., Bruzzone, L. (2004) Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on geoscience and remote sensing*, 42, 1778–1790.

[2] Zullo, A., Fauvel, M., Ferraty, F., Goulard, M., Vieu, P. (2014) Non-parametric functional methods for hyperspectral image classification. *In Geoscience and Remote Sensing Symposium (IGARSS), 2014 IEEE International*, 3422–3425.

[3] Ordóñez, C., Oviedo de la Fuente, M. O., Roca, J., & Rodríguez, J.R. (2017). Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach. *Chemometrics and Intelligent Lab. Sys.*

[4] Febrero-Bande, M., González-Manteiga, W., Oviedo de la Fuente, M. O. (2017). Variable selection in Functional Additive Regression Models.*In Functional Statistics and Related Fields (pp. 113-122).* Springer, Cham.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

UN TEST ESTADÍSTICO PARA ANALIZAR LA VARIABILIDAD ESPACIAL DE LOS DATOS USANDO COMPONENTES PRINCIPALES GEOGRÁFICAMENTE PONDERADAS

J. Roca-Pardiñas[1], C. Ordóñez[2]

[1]Departamento de Estadística e Investigación Operativa, Universidad de Vigo, 36310 Vigo, España
[2]Departamento de Explotación y Prospección de Minas, Universidad de Oviedo, 33004 Oviedo, España

RESUMEN

En este trabajo se propone un método para evaluar la variabilidad espacial de los datos a partir de la estructura de la matriz de covarianzas en un análisis de componentes principales geográficamente ponderado (GWPCA). El método se basa en realizar un test de hipótesis sobre los autovectores de las puntuaciones de las componentes principales en vez de utilizar el método habitual basado en calcular los autovalores de la matriz de covarianzas. Nuestra propuesta se evaluó con datos simulados, y se comprobó que tiene mayor potencia estadística que el método habitual. Finalmente se aplicó a un problema con datos reales cuyo objetivo es encontrar patrones de distribución espacial en un conjunto de elementos contaminantes de suelos. El resultado obtenido muestra la utilidad de GWPCA frente al análisis clásico de componentes principales (PCA).

Keywords and key sentences: variabilidad espacial, suavizado de núcleo, ancho de banda, componentes principales, contaminación de suelos.

1. INTRODUCCIÓN

El análisis de componentes principales geográficamente ponderado (GWPCA) es una extensión del análisis de componentes principales clásico (PCA) para situaciones en las que se asume que la estrucutura de la matriz de covarianzas no es constante en el espacio [1]. En esencia GWPCA realiza un análisis de componentes principales para cada observación en un entorno del espacio alrededor de dicha observación en el cual se asume que existe homogeneidad en la covarianza. La determinación del tamaño de esa vecindad (ancho de banda) es uno de los aspectos esenciales de GWPCA. Cuando el tamaño de banda es suficientemente grande GWPCA se reduce a PCA ([2]). Aunque resultados del GWPCA como las cargas o el porcentaje de varianza explicado por cada componente muestre variabilidad espacial, un análisis riguroso debe ir acompañado de un análisis estadístico que pruebe que dicha variabilidad es significativa estadísticamente, de otra manera el uso de GWPCA frente a PCA

no estaría justificado. En [3] se propuso un test de Monte Carlo para analizar la signifación de la variabilidad de los valores propios cuya idea es determinar la desviación típica de cada valor propio local en un rango de distribución de las desviaciones típicas obtenido aplicando GWPCA para cada conjunto de datos aleatorio. Este test está implementado en el paquete de R GWmodel [4]. En este artículo presentamos un método diferente basado en definir un estadístico que usa los autovectores de la matriz de covarianzas y estimar su nivel de significación a partir de la función de distribución del estadístico obtenida también mediante simulación de Monte Carlo.

2. METODOLOGÍA

Para estudiar la variabilidad espacial de los datos que justifique el uso de GWPCA desde un punto de vista estadístico se propone el siguiente test de hipótesis

$$H_0 : \Sigma(\mathbf{s}) = \Sigma \quad \forall \mathbf{s} \quad \text{frente a} \quad H_1 : \Sigma(\mathbf{s}) \text{ no siempre iguales} \quad \forall \mathbf{s} \tag{1}$$

donde $\Sigma(\mathbf{s})$ representa la matriz de covarianzas dependiente de la posición espacial $\mathbf{s} = (\mathbf{s_1}, \mathbf{s_2})$
En [3] se propuso el siguiente estadístico

$$T_1 = n^{-1} \sum_{i=1}^{n} \left(\hat{\lambda}_1(\mathbf{s}_i) - \bar{\lambda}_1 \right)^2 \tag{2}$$

donde $\hat{\lambda}_1(\mathbf{s}_i)$ es el primer valor propio de la matriz y $\bar{\lambda} = n^{-1} \sum_{i=1}^{n} \hat{\lambda}_1(\mathbf{s}_i)$. Obviamente, si no existe variabilidad espacial el valor de T_1 debe ser nulo.
En este trabajo proponemos otro estadístico distinto definido de la siguiente manera

$$T_2 = n^{-1} \sum_{i=1}^{n} \left(\hat{\mathbf{P}}_{\mathbf{z}}^{(1,1)}(\mathbf{s}_i) - 1 \right)^2 \tag{3}$$

donde $\hat{\mathbf{P}}_{\mathbf{z}}^{(1,1)}$ representa el elemento $(1,1)$ de la matriz $\mathbf{P_z}$ que proviene de la descomposición espectral de la matriz de covarianzas $\Sigma_{\mathbf{z}}(\mathbf{s}) = \mathbf{P_z}(\mathbf{s})\Lambda_{\mathbf{z}}(\mathbf{s})\mathbf{P}_{\mathbf{z}}^t(\mathbf{s})$.
Si no existe estructura espacial en los datos, la primera columna de será próxima a $(1, 0..., 0)$, la segunda a $(0, 1, .., 0)$, y así sucesivamente. Entonces, el estadístico T_2 también debe ser próximo a 0. Para un nivel de significancia α dado la regla de decisión consiste en rechazar la hipótesis nula si $T > \hat{T}^{\alpha}$, donde $\hat{T}^{\alpha}$ representa el percentil empírico $(1-\alpha)$ de los valores $T^{*1}, \ldots, T^{*B}$. El procedimiento seguido es el siguiente:
Para $b = 1, \ldots, B$ (por ejemplo, $B = 1000$),
Paso 1: Obtener la muestra $\mathbf{s}_1^{*,b}, \ldots, \mathbf{s}_n^{*,b}$ mediante permutación de los datos originales $\mathbf{s}_1, \ldots, \mathbf{s}_n$.
Paso 2: Calcular el test estadístico T^{*b} usando los datos originales $\mathbf{x}_1, \ldots, \mathbf{x}_n$ y los remuestreados $\mathbf{s}_1^{*,b}, \ldots, \mathbf{s}_n^{*,b}$

3. EVALUACIÓN SOBRE DATOS SIMULADOS

Para evaluar el comportamiento del método propuesto se compararon los test estadístico T_1 y T_2 en una muestra de mil observaciones $\{\mathbf{s}_i, \mathbf{x}_i\}_{i=1}^{n}$ extraídas de una distribución uniforme bivariante $U[-2,2] \times U[-2,2]$, donde $\mathbf{x}_i$ es un vector $p-$dimensional que tiene una distribución gaussiana de media cero y matriz de covarianzas

$$\Sigma(\mathbf{s}_i) = \begin{pmatrix} 1 & \rho(\mathbf{s}_i) & \ldots & \rho(\mathbf{s}_i) \\ \rho(\mathbf{s}_i) & 1 & \ldots & \rho(\mathbf{s}_i) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(\mathbf{s}_i) & \rho(\mathbf{s}_i) & \ldots & 1 \end{pmatrix} \tag{4}$$

Se consideró que $\rho(\mathbf{s}_i) = \min\left(a\left|s_{i1}^2 - s_{i2}^2\right|, 0.99\right)$, por lo que para $a = 0$ se debe cumplir la hipótesis nula, mientras que a medida que el valor de a crece el modelo se aparta de dicha hipótesis. Las curvas de potencia de T_1 y T_2 se muestran en la Figura 1. Como se puede observar, el estadístico T_2 tiene un mejor comportamiento que T_1.

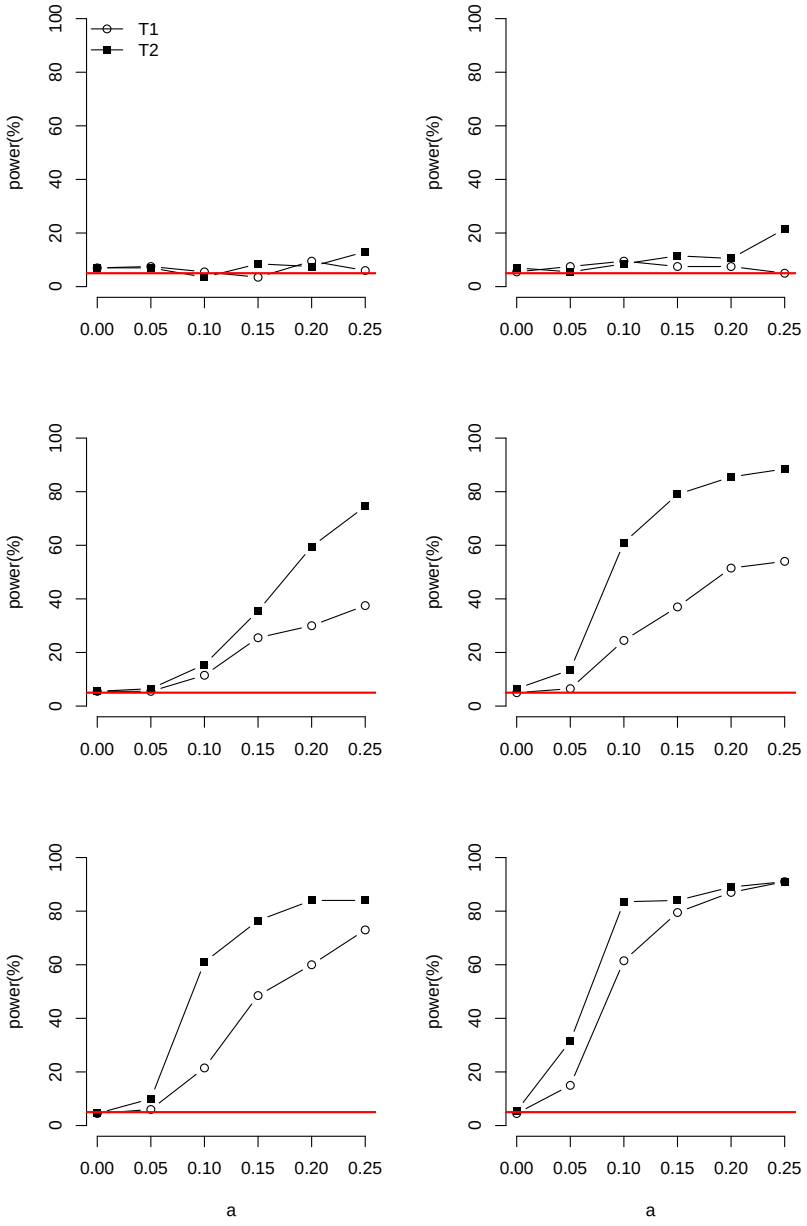


Figure 1: Porcentaje de rechazos para los estadísticos T_1 y T_2 en función de a, para un nivel de significación del 5% y un número de covariables $p = 5$ and $p = 10$ (paneles izquierdo y derecho, respectivamente). Panel superior: rechazos para un tamaño de muestra $n = 100$. Panel medio: rechazos para $n = 500$. Panel inferior: rechazos para $n = 1000$.

4. CASO DE ESTUDIO CON DATOS REALES

El método propuesto se utilizó para analizar el patrón de variabilidad de una serie de contaminantes de suelo en una zona industrial localizada en el noroeste de España. Se calcularon los $p-$valores obtenidos para ambos estadísticos en función del tamaño de la vecindad (h) y

se determinó que no hay significación estadística para valores de h inferiores a 0.30 (30% of the data set). En concreto, se seleccionó un ancho de banda de 0.28 para llevar a cabo el GWPCA, obtenido mediante validación cruzada. La Figura 2 muestra la distribución espacial del porcentaje de varianza explicada por las cuatro primeras componentes principales. El hecho de que existan cambios en dicha variabilidad espacial, confirma la idoneidad de aplicar GWPCA en vez de PCA.

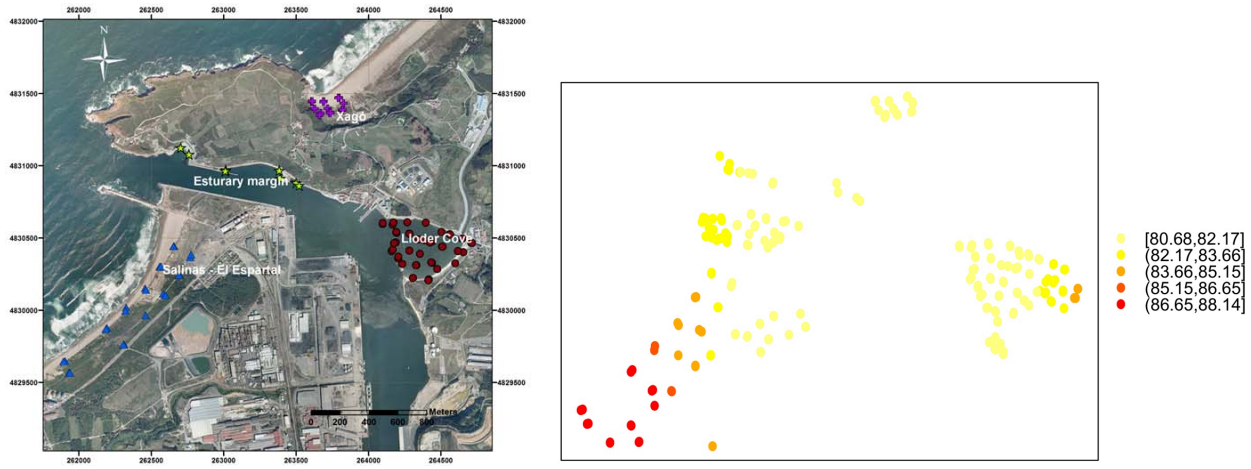


Figure 2: Localización espacial de las muestras (izquierda) y porcentaje de varianza local para las primeras cuatro componentes principales (derecha).

5. CONCLUSIONES

La principal contribución de este trabajo es la propuesta de un test estadístico para contrastar la existencia de variabilidad espacial que se necesita para justificar el uso de GWPCA en vez de PCA. De acuerdo con el estudio de simulación realizado, el test que proponemos produce una potencia estadística superior a la de otro test diseñado anteriormente por otros autores. Por tanto, nuestro test tiene menos probabilidad de considerar que existe homogeneidad espacial en los datos cuando realmente no la hay. Sin embargo, hemos constatado que los resultados obtenidos dependen del ancho de banda, esto es, del tamaño de la vecindad elegido. El problema de la elección del ancho de banda es complejo, siendo una posible alternativa utilizar validación cruzada para realizar una estimación inicial. El conocimiento de cada problema en particular puede ayudar a modificar en la dirección adecuada dicho valor inicial.

References

[1] Tipping,M.E., and Bishop, Ch.M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B* 61(3), 611–622.

[2] Demsar, U., Harris, P., Brunsdon, Ch., Fotheringham, A., and McLoone A. (2013). Principal Component Analysis on Spatial Data: An Overview. *Annals of the Association of American Geographers* 103 (1), 106–128.

[3] Harris P., Brunsdon C., and Charlton M. (2011). Geographically weighted principal component analysis. *International Journal of Geographical Information Science* 25(10), 1717–1736.

[4] Gollini, I., Lu, B., Charlton, M., Brunsdon, Ch., and Harris, P. (2015). GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models 2015. *Journal of Statistical Software* 63(17), 1–50.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

A INFLUÊNCIA DA GESTÃO NA PRODUTIVIDADE DE PLANTAÇÕES DE *EUCALIPTUS GLOBULUS*

Catarina Monteiro[1], Nélia Silva[2], Isabel Pereira[3]

[1] Departamento de Matemática da Universidade de Aveiro
[2] Departamento de Matemática da Universidade de Aveiro e CIDMA
[3] Departamento de Matemática da Universidade de Aveiro e CIDMA

RESUMO

No cerne da atualidade, o ambiente, a sustentabilidade e a gestão florestal ganham um destaque particular, assumindo um papel preponderante em inúmeras unidades da economia, mormente, no setor florestal. Inerente ao setor florestal evidencia-se o destaque como o setor da maior importância para Portugal, sendo dos poucos cuja atividade promove os três grandes pilares da sustentabilidade: económico, social e ambiental. A nível económico, salienta-se a indústria da pasta, papel e cartão, a qual possui relevante interesse estratégico para Portugal, na economia Portuguesa, com especial realce para o mercado de exportações, sendo um dos maiores exportadores de valor acrescentado nacional.

Considerando os crescentes desafios que Portugal enfrenta, inúmeras entidades de destaque, consideram que o papel produzido através do setor florestal pode ser incrementado quantitativa e qualitativamente, assumindo um papel cada vez mais relevante na economia nacional, contribuindo fortemente para o crescimento e desenvolvimento económico do País. Sendo Portugal um país pautado por condições de solo e clima de excelência que permitem, no quadro europeu, um desenvolvimento florestal competitivo, a que acrescem a disponibilidade de área territorial, é primordial o desenvolvimento de técnicas de gestão florestal que maximizem a produtividade. Permitindo assim evidenciar o potencial de crescimento deste setor nos três pilares de sustentabilidade supracitados. Alguns empresários e gestores de microempresas questionam frequentemente a importância e finalidade da gestão florestal. Há, inclusive, quem desconheça a possibilidade de gerir as suas parcelas por meio de tecnologias de última geração, e consultores especializados em planeamento, operações e produtividade. Desta forma, na presente investigação pretende-se evidenciar os benefícios económicos da gestão florestal em áreas de minifúndio florestal, mormente na otimização da Produtividade das referidas áreas.

Foi desenvolvida uma análise preliminar dos dados, que permitiu constatar a existência de uma panóplia de variáveis indicativas da Produtividade da Parcela, sobre as quais se aplicou Análise em Componentes Principais (ACP), tendo sido obtida uma só variável Produtividade, sobre a qual os estudos subsequentes incidem. Objetivando-se intuir quanto às vantagens da aplicação de diferentes metodologias de gestão florestal sob a Produtividade das parcelas, foram

formados cinco grupos com base na existência/inexistência de gestão e na entidade responsável pela parcela. Assim, fez-se a comparação da produtividade entre os grupos de parcelas estabelecidos com recurso a metodologias estatísticas não paramétricas adequadas (incluindo a ANOVA não paramétrica a dois fatores), e a correspondente implementação no software R.

Palavras e frases chave: *Eucaliptus globulus*, Análise em Componentes Principais, Gestão florestal, Produtividade, ANOVA *two-way* não paramétrica.

Referências

[1] Maroco, J. (2010). Análise Estatística com utilização do SPSS. (Sílabo, Ed.) (3a Edição.).

[2] Murteira, B., Silva Ribeiro, C., Andrade e Silva, J. & Pimenta, C. (2010). Introdução à Estatística. (Escolar Editora, Ed.).

[3] Guimarães, R. C., & Cabral, J. S. (1997). Estatística. (Mc Graw Hill, Ed.).

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

ENTROPIA NORMALIZADA E OUTROS MÉTODOS DE SELEÇÃO DE VARIÁVEIS: UM ESTUDO COMPARATIVO COM DADOS SIMULADOS

Alberto Oliveira da Silva[1], Rodney Sousa[1] e Pedro Macedo[1]

[1] CIDMA, Departamento de Matemática, Universidade de Aveiro

RESUMO

Este trabalho examinou o problema da seleção de variáveis em modelos de regressão linear. Foram considerados dois modelos com dados simulados e três métodos de seleção: o LASSO, o *Elastic Net* e a Entropia Normalizada (NE). Duas estratégias para a definição dos suportes no método da máxima entropia generalizada (GME) foram consideradas, nomeadamente as estimativas de regressão *ridge* e as estimativas do método *Partial Least Squares* (PLS). Verificou-se um melhor desempenho da NE, com uma perfeita identificação das variáveis relevantes, independentemente da opção usada na especificação dos suportes para o GME. Todavia, constatou-se uma elevada contração nos valores das estimativas obtidas pelo GME, ao usarem-se as estimativas de regressão *ridge* na definição dos suportes, sendo menor a contração com a utilização das estimativas do PLS na definição dos suportes para o GME. Verificou-se, adicionalmente, que o LASSO e o *Elastic Net* tendem a estimar modelos mais complexos.

Palavras e frases chave: *Elastic Net*; Entropia Normalizada; LASSO; PLS; Traço *Ridge*.

1. INTRODUÇÃO

Devido ao desenvolvimento tecnológico observado nas últimas décadas, é comum que, nos dias de hoje, bases de dados com origem em várias áreas das Ciências da Vida possuam muitas variáveis. Em um contexto de regressão linear, essa profusão de variáveis pode tornar-se um desafio, especialmente quando se tratam de problemas mal condicionados. Considerando-se que a matriz de dados que representa as variáveis explicativas de muitos problemas reais (*e.g.*, da Bioquímica) pode conter, eventualmente, colunas fortemente correlacionadas, muitos problemas de estimação de modelos de regressão podem ser caracterizados como mal condicionados, pelo que podem apresentar soluções altamente instáveis. Ao lidar com grandes volumes de dados, o melhor modelo nem sempre é o que contém mais variáveis, uma vez que tais modelos podem resultar em sobrestimação (*overfitting*). Portanto, é importante que um método de estimação

seja capaz de considerar apenas as variáveis relevantes para o melhor ajuste do modelo ao conjunto de dados, assim como de ter qualidade preditiva, avaliada segundo algum critério de aferição (*e.g.*, baseado no menor erro quadrático médio, MSE). Visando atingir tais objetivos, têm sido propostos diversos métodos de regularização e de redução da dimensionalidade, tais como o LASSO (Tibshirani, 1996), o OSCAR (Bondell e Reich, 2007), o *Elastic Net* (Zou and Hastie, 2005) e a NE (Golan *et al.*, 1996). Os primeiros são métodos muito populares entre os investigadores, mas a NE é pouco divulgada e parece ser uma boa alternativa. Com o objetivo de comparar o desempenho de alguns dos métodos usuais, este trabalho considera duas matrizes de preditores simuladas, com números de condição distintos, 1 e 200, ambas com 30 observações e 20 variáveis explicativas.

2. ALGUNS RESULTADOS

O principal desafio com a utilização da NE é a escolha adequada dos suportes para o estimador GME. Diversos estudos de simulação revelam que a utilização de suportes de grande amplitude no GME, simétricos e igualmente espaçados em torno de zero, conduzem a valores de NE próximos de um, pelo que a seleção de variáveis deixa de ser possível. Este trabalho propõe duas abordagens para a definição dos limites dos suportes: uma escolha baseada nas estimativas máximas obtidas pela regressão *ridge*, ilustradas e visualizadas no traço *ridge* para um intervalo adequado do parâmetro de contração (ridGME) e uma escolha baseada nas estimativas obtidas pelo PLS (plsGME). Os resultados da seleção de variáveis são apresentados nas Tabelas 1-3, sendo, naturalmente, conhecidas as três variáveis relevantes: X_2, X_5 e X_{18}.

Seleção	**LASSO$_{\lambda=0.248}$**	**Elastic Net$_{(\lambda=0.268,\ \alpha=0.7)}$**	**NE - ridGME**	**NE - plsGME**
$\boldsymbol{X_2}$	✓	✓	✓	✓
$\boldsymbol{X_5}$	✓	✓	✓	✓
$\boldsymbol{X_{18}}$	✓	✓	✓	✓
Outras	X_{11}	X_3, X_6, X_{11}		

Tabela 1: Variáveis selecionadas – matriz com número de condição 1.

Seleção	**LASSO$_{\lambda=0.2341}$**	**Elastic Net$_{(\lambda=0.169,\ \alpha=0.7)}$**	**NE - ridGME**	**NE - plsGME**
$\boldsymbol{X_2}$	✓	✓	✓	✓
$\boldsymbol{X_5}$	✓	✓	✓	✓
$\boldsymbol{X_{18}}$	✓	✓	✓	✓
Outras	X_7, X_{17}	X_7, X_{15}, X_{17}		

Tabela 2: Variáveis selecionadas – matriz com número de condição 200.

Estimativas	LASSO	Elastic net	NE - ridGME	NE - plsGME	PLS
$\widehat{\boldsymbol{\beta}}_{\mathbf{2}} = \mathbf{12}$	10.15	10.20	2.43	4.84	7.95
$\widehat{\boldsymbol{\beta}}_{\mathbf{5}} = \mathbf{8}$	5.86	5.89	2.17	2.27	8.28
$\widehat{\boldsymbol{\beta}}_{\mathbf{18}} = -\mathbf{15}$	-13.62	-13.61	-3.17	-7.8	-14.53
$\widehat{\boldsymbol{\beta}}_{\mathbf{7}} = \mathbf{0}$	0.15	0.25			
$\widehat{\boldsymbol{\beta}}_{\mathbf{15}} = \mathbf{0}$		0.07			
$\widehat{\boldsymbol{\beta}}_{\mathbf{17}} = \mathbf{0}$	0.43	0.54			

Tabela 3: Estimativas obtidas pelos métodos utilizados – matriz com número de condição 200.

3. CONCLUSÕES

Os resultados mostram que a NE tem um desempenho superior ao dos métodos LASSO e *Elastic Net* na correta identificação das variáveis relevantes, independentemente do bom ou mau condicionamento da matriz dos regressores. As escolhas dos limites (superior e inferior) dos suportes para o GME, baseadas nas estimativas máximas obtidas pela regressão *ridge* (num dado intervalo admissível para o parâmetro de contração) e nas estimativas obtidas pelo PLS, foram adequadas para a correta identificação das variáveis relevantes através da NE. Tendo-se verificado, no entanto, que a escolha dos suportes por intermédio do traço *ridge* conduz a elevada contração das estimativas obtidas pelo GME, duas alternativas revelaram-se promissoras para mitigar o problema: (a) utilizar as estimativas do método PLS para a definição dos suportes, tendo, como resultado, uma contração mais branda realizada pelo GME; (b) utilizar as estimativas do PLS para definir os suportes para o GME e, selecionadas as variáveis pela NE, considerar as estimativas correspondentes obtidas pelo PLS, obtendo-se estimativas mais próximas dos valores reais dos parâmetros. Por último, e como se trata de um trabalho preliminar, sugere-se que, futuramente, seja desenvolvido um estudo que estabeleça as relações teóricas entre a NE e o PLS, dado que esta relação alcança, aparentemente, bons resultados, até mesmo em relação ao LASSO e ao *Elastic Net* em termos de contração, além de permitir a sua implementação em problemas com mais preditores do que observações. Adicionalmente, como a seleção dos suportes para o GME, baseada em estimativas de outras metodologias, é uma escolha claramente subjetiva, um amplo estudo de simulação deverá ser implementado para aferir a sensibilidade da NE a estas e a outras possíveis escolhas de suportes.

AGRADECIMENTOS

Trabalho parcialmente suportado por fundos portugueses através do CIDMA (Centro de Investigação e Desenvolvimento em Matemática e Aplicações), da Universidade de Aveiro, e da FCT (Fundação para a Ciência e a Tecnologia), através do projeto UID/MAT/04106/2013.

Referências

[1] Bondell H., Reich B. (2007). Regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* 64, 115-123.

[2] Golan, A., Judge, G., Miller, D. (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. Wiley, Chichester.

[3] Tibshirani R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B* 58, 267-288.

[4] Zou H., Hastie T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B* 67, 301-320.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

A LACK-OF-FIT TEST FOR QUANTILE REGRESSION MODELS USING LOGISTIC REGRESSION

Mercedes Conde-Amboage[1], Valentin Patilea[2] and César Sánchez-Sellero[1]

[1]Department of Statistics, Mathematical Analysis and Optimization. University of Santiago de Compostela. Spain.
[2]Center of Research in Economics and Statistics (Ensai). France.

ABSTRACT

A new lack-of-fit test for parametric quantile regression models is proposed. The test is based on interpreting the residuals from the quantile regression model fit as response values of a logistic regression, the predictors of the logistic regression being functions of the covariate of the quantile model. Then a correct quantile model implies the nullity of all the coefficients but the constant in the logistic model. Given this property, we use a lack-of-fit test in the logistic regression to check the quantile regression model. When the functions of the covariate are taken from a basis and for a large enough number of functions, our test can detect general departures from the parametric quantile model. To approximate the critical values of the test, a wild bootstrap mechanism is used.The good properties of the new test versus other nonparametric tests available in the literature are shown by means of a simulation study. A real data application is given to illustrate the new methodology.

Keywords and key sentences: quantile regression, lack-of-fit testing, logistic regression.

1. INTRODUCTION

Given a pair of variables (X, Y), where Y is the response variable and X is an explanatory variable, let us consider a quantile regression model denoted by

$$Y = q_\tau(X) + \varepsilon,$$

where $q_\tau(\cdot)$ represents the regression function and the error ε has a conditional τ-quantile equal to zero, that is $P(\varepsilon \leq 0 | X = x) = \tau$ for almost all x. See [5] for a seminal paper on quantile regression. Along this paper, we will focus on the problem of testing a parametric quantile regression model

$$H_0 : q_\tau(\cdot) \in \mathcal{Q}_\theta = \{q_\tau(\cdot, \theta) \ : \ \theta \in \Theta \subset R^q\}, \tag{1}$$

versus a nonparametric alternative. This problem was addressed by [3], [7] and [4], among others.

The new lack-of-fit test will be based on an idea introduced by [6]. They have proposed a simple method using logistic regression to identify significant covariates associated with a quantile regression model. Their development relies on the fact that observations can be classified as above or below the predicted quantile. This classification step creates a dichotomous variable that can be utilized as the response variable in a logistic regression model. If the probability of being above the predicted quantile is independent of a certain explanatory variable, then this probability will be a constant across all values of the explanatory variable indicating no association between the quantile and the explanatory variable. Otherwise, if an explanatory variable within logistic regression is statistically significant, the same variable is interpreted to be significant in the quantile regression model.
Here we will extend this parametric significance to a nonparametric significance, in order that any kind of deviation from the parametric model can be detected. To do so, we apply an orthogonal series fit as explained and studied by [1].

2. THE PROPOSED METHOD

The dichotomous variable associated with the error of a parametric quantile regression model is defined as

$$Z(\theta) = I(Y \leq q_\tau(X, \theta)),$$

where $I(\cdot)$ is the indicator function of an event. Then, the parametric quantile regression model is correct if and only if there exists some $\theta \in \Theta$ such that the conditional probability of $Z(\theta)$ given X does not depend on X, and is equal to τ. At this point, in order to check the independence between a suitable $Z(\theta)$ and X, the idea is to consider a logistic regression with response $Z(\hat{\theta})$, where $\hat{\theta}$ is an estimator of θ, and many covariates obtained as functions of the components of the vector X, and to test the nullity of all the coefficients but the constant.
We are going to focus on a univariate X. In this situation, to detect general nonparametric alternatives, the vector W used as explanatory variable in the logistic regression should contain as many functions of the components of X, as needed in order to detect all kind of alternative hypothesis. To formally describe our procedure, these function are represented by a dense basis of functions. Although different basis of functions can be considered, we will make use of Hermie polynomials, defined by:

$$H_p(x) = p! \sum_{m=0}^{[p/2]} \frac{(-1)^m}{m!(p-2m)!} \frac{x^{p-2m}}{2^m}, \qquad x \in R, \qquad p \in \{0, 1, 2, \ldots\},$$

where $[a]$ denotes the integer part of a real number a.
Then, the idea is to check whether, for some value θ, we have $\varphi_1 = \varphi_2 = \cdots = \varphi_p = 0$ in the logistic regression model:

$$\text{logit}\left(P[Z(\theta) = 1|X]\right) = \varphi_0 + \varphi_1 H_1(X) + \varphi_2 H_2(X) + \cdots + \varphi_p H_p(X) = \varphi' W,$$

where $\varphi = (\varphi_0, \varphi_1, \ldots, \varphi_p)$ is a vector of coefficients.
Given a sample of independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, the estimator $\hat{\theta}$ is obtained as the minimizer of

$$\sum_{i=1}^{n} \rho_\tau \left(Y_i - q_\tau(X_i, \theta)\right),$$

where $\rho_\tau(u) = u\big(\tau - I(u < 0)\big)$ is the well-known quantile loss function. Then, the responses for the logistic regression model are constructed as $Z_i(\hat{\theta}) = I(Y_i \leq q_\tau(X, \hat{\theta}))$, while the vectors of explanatory variables are computed as $W_i = (1, H_1(X_i), \ldots, H_p(X_i))$, for $i \in \{1, \ldots, n\}$.

The log-likelihood function in the logistic regression model is given by

$$L_p(\varphi) = \sum_{i=1}^{n} \Big(Z_i(\hat{\theta}) \varphi' W_i - \log(1 + e^{\varphi' W_i}) \Big)$$

To check the significance of the coefficients φ but φ_0, we use the lack-of-fit test for logistic regression models presented by [1]. Let us denote $U_p(\varphi)$ the column vector of first derivatives of L_p with respect to φ and denote $A_p(\varphi)$ to the expected information matrix. The following score statistic is considered

$$S_p = U_p(\hat{\varphi}_{p,0})^T \; A_p(\hat{\varphi}_{p,0})^{-1} \; U_p(\hat{\varphi}_{p,0})$$

where $\hat{\varphi}_{p,0} = (\hat{\varphi}_{\text{H0},0}, 0, \ldots, 0)$, $\hat{\varphi}_{\text{H0},0}$ being the estimator of φ_0 under the null hypothesis of no-effect.
The following test statistics will be considered to test the lack-of-fit of the quantile regression model:

$$T_1 = S_{\hat{p}} \qquad T_2 = \frac{S_{\hat{p}} - \hat{p}}{\max\{1, \sqrt{\hat{p}}\}} \qquad T_3 = \text{SIC}(\hat{p}, 2)$$

where $\text{SIC}(p, c) = S_p - c \cdot p$ with $r = 0, 1, \ldots$, and $\hat{p} = \arg\ \max_p\ \text{SIC}(p, 2)$.
A bootstrap procedure adapted to the quantile regression context is proposed in order to calibrate the critical values of the given test statistics. The bootstrap procedure works as follows:

1. Let $\varepsilon_i^\star = \delta_i |r_i|$, where $r_i = Y_i - q_\tau(X_i, \hat{\theta}_\tau)$ are the residuals from the original sample. The multipliers, δ_i, are independently generated from the two-point distribution with probabilities $(1 - \tau)$ and τ at $2(1 - \tau)$ and -2τ, respectively (see [2]). Compute $Y_i^\star = q_\tau(X_i, \hat{\theta}_\tau) + \varepsilon_i^\star$ for each $i = 1, \ldots, n$.

2. Use the bootstrap data set $\{(X_i, Y_i^\star),\ i = 1, \cdots, n\}$ to compute the bootstrap estimator $\hat{\theta}_\tau^\star$ and the dichotomous variables $Z_i(\hat{\theta}_\tau^\star) = I(Y_i^\star \leq q_\tau(X_i, \hat{\theta}_\tau^\star))$.

3. Use the data set $\{(W_i, Z_i(\hat{\theta}_\tau^\star)),\ i = 1, \cdots, n\}$ to compute the bootstrap test statistics T_1^*, T_2^* and T_3^*.

4. Repeat Steps 1, 2 and 3 B times, and estimate the α-level critical values by the $(1 - \alpha)$-quantile of the resulting B values of each statistic.

3. SIMULATION STUDY

The performance of the proposed method under the null and alternative hypotheses will be analysed using a Monte Carlo simulation study. The number of simulated original samples was 1000 and the number of bootstrap replications was 500.
Here we will only show the results of a power comparison under the following model:

$$\text{Model 1: } Y = 1 + X + h(X) + \varepsilon,$$

where X and ε are independent and follow a standard Gaussian distribution. We are going to test the linearity of the conditional median. Then, the function h represents the deviation from the null hypothesis. The following deviations will be considered: D1. $h(x) = \frac{1}{2}x^2$; D2. $h(x) = \frac{1}{2}\sin(\pi x)$ and D3. $h(x) = 4\phi(x)$, where ϕ represents the density function associated with a standard Gaussian distribution.
Percentages of rejections are given in Table 1 for the three statistics proposed here and for other three competitors, when they are applied with a nominal level of 5%. Although no test is better than the others in all cases, we can conclude that the test based on the statistic T_3 shows a globally better performance than the other tests.

	n	T_1	T_2	T_3	Z	HZ	HS
D1	50	16.8	61.8	67.7	25.0	65.7	71.5
	100	66.9	95.2	96.6	55.1	96.3	96.6
	200	98.4	99.9	100.0	92.6	100.0	99.9
D2	50	22.8	21.3	19.3	16.3	7.5	8.9
	100	52.5	43.5	35.7	38.9	7.6	8.7
	200	83.7	79.9	73.9	74.0	17.5	16.3
D3	50	14.1	47.9	53.8	19.8	45.0	35.3
	100	51.5	83.2	88.5	45.5	83.8	81.6
	200	90.4	99.1	99.6	86.8	99.9	99.2

Table 1: Percentages of rejections associated with the new test (T_1, T_2 and T_3), the test proposed by Zheng (1998) (Z), the test proposed by He and Zhu (2003) (HZ) and the test proposed by Horowitz and Spokoiny (2002) (HS) under Model 1 and deviations D1, D2 and D3.

3. A REAL DATA APPLICATION

We have applied the proposed methodology to an environmental dataset, *airquality*, which is available in the basic installation of $\mathcal{R}$. This dataset consists of daily air quality measurements in New York, from May 1 to September 30, 1973. We will focus on the variables: Ozone, which contains mean ozone concentrations measured from 13:00 to 15:00 hours at Roosevelt Island, and Temp, which contains maximum daily temperature measured at La Guardia Airport. The three test statistics given here, T_1, T_2 and T_3, provide very low p-values for the tests of linearity in the regression effect of Temperature on the variable Ozone.

References

[1] Aerts, M., Claeskens, G., Hart, J.D. (2000). Testing lack of fit in multiple regression. *Biometrika*, 87, 405–424.

[2] Feng, X., He, X., Hu, J. (2011). Wild bootstrap for quantile regression. *Biometrika*, 98, 995–999.

[3] He, X., Zhu, L.-X. (2003). A lack-of-fit test for quantile regression. *Journal of the American Statistical Association*, 98, 1013–1022.

[4] Horowiz, J.L., Spokoiny, V.G. (2002). An adaptive, rate-optimal test of linearity for median regression models. *Journal of the American Statistical Association*, 97, 822–835.

[5] Koenker, R., Bassett, G. (1978). Regression quantiles, *Econometrica*, 46, 33–50.

[6] Redden, D.T., Fernández, J.R., Allison, D.B. (2004). A simple significance test for quantile regression. *Statistics in Medicine*, 23, 2587–2597.

[7] Zheng, J.X. (1998). A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econometric Theory*, 14, 123–138.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

METODOLOGIAS DE MÁXIMA ENTROPIA NA ANÁLISE DE DADOS EM LARGA ESCALA

Maria da Conceição Costa [1] e Pedro Macedo[1]

[1]CIDMA, Departamento de Matemática, Universidade de Aveiro

RESUMO

A relação entre teoria da informação, estatística e máxima entropia foi estabelecida na década de cinquenta do século passado, na sequência dos trabalhos de Kullback, Leibler, Lindley e Jaynes. O domínio de aplicação abrangia, essencialmente, a área das telecomunicações, mas, inevitavelmente, outros domínios científicos, tais como a medicina, a biologia, a economia ou as engenharias, tornaram-se rapidamente grandes beneficiários desta profícua relação. O interesse crescente na convergência entre métodos de análise estatística, processamento de informação, ciências computacionais e teoria da decisão deu origem a uma nova área de investigação, designada por *Info-Metrics*, que se assume como um conjunto de metodologias de optimização, com vista ao processamento de informação, modelação e inferência em problemas mal-postos, ou seja, na presença de problemas com informação finita, incompleta e com enorme ruído, entre outras características indesejáveis; e.g., [**?**], [**?**]. Todavia, embora adequada a este tipo de problemas, a abrangência desta área não se confina aos problemas mal-postos, podendo fornecer uma visão complementar a algumas metodologias estatísticas tradicionais.

A crescente disponibilidade e utilização de dados em larga escala condiciona, frequentemente, a utilização de algumas metodologias estatísticas tradicionais. Muitas observações, muitas variáveis, dados recolhidos em diferentes regimes temporais ou a partir de múltiplas fontes são características que, comummente, originam enormes desafios à análise estatística. Um aspecto particular que este trabalho aborda, e que é uma consequência habitual das propriedades referidas, é a presença de heterogeneidade nos dados, que impossibilita a aplicação das técnicas clássicas de regressão. As abordagens estatísticas tradicionais, capazes de lidar com a questão da falta de homogeneidade, apresentam a desvantagem de ter uma carga computacional demasiado elevada para dados em larga escala. Recentemente, foram desenvolvidas abordagens alternativas baseadas em metodologias de agregação de dados; e.g., [**?**], [**?**]. Neste trabalho, é proposta uma nova abordagem ao problema da análise de dados heterogéneos em larga escala, baseada na metodologia *Info-Metrics*, onde o conceito de máxima entropia normalizada é introduzido nos procedimentos de agregação. O bom desempenho desta nova abordagem metodológica é ilustrado através de um estudo de simulação e confirmado na aplicação a dois conjuntos de dados reais provenientes da área do controlo de qualidade.

Palavras-chave: dados em larga escala, *info-metrics*, máxima entropia.

AGRADECIMENTOS

Este trabalho é suportado pelo CIDMA – Centro de Investigação e Desenvolvimento em Matemática e Aplicações e pela FCT – Fundação para a Ciência e a Tecnologia, no âmbito do projecto UID/MAT/04106/2013.

Referências

[1] A. Golan (2013). On the State of Art of Info-Metrics, in *Uncertainty Analysis in Econometrics with Applications*. Van Nam Huynh, Vladik Kreinovich, Songsak Sriboonchitta, and Komsan Suriya, Springer-Verlag, Berlin, pp. 3–15.

[2] A. Golan (2018). *Foundations of Info-Metrics, Modeling, Inference, and Imperfect Information*. Oxford University Press, New York.

[3] N. Meinshausen, and P. Bühlmann (2015). Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4), 1801–1830.

[4] P. Bühlmann, and N. Meinshausen (2014). Magging: maximin aggregation for inhomogeneous large-scale data. *arXiv:1409.2638v1*.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

ANÁLISE EXPLORATORIA DA DISTRIBUCIÓN ESPACIAL DOS CENTENARIOS DA GALIZA

Carlos L. Iglesias Patiño[1] e M. Esther López Vizcaíno[1]

[1] Instituto Galego de Estatística

RESUMO

Abordamos a distribución espacial dos habitantes centenarios de Galicia no 2016 co obxectivo de coñecermos este colectivo, vermos se se pode supor aleatoria e validarmos as fontes e, nomeadamente, o proceso de xeorreferenciación realizado, mediante técnicas baseadas nunha Poisson ou sen considerarmos distribución.

Palabras e frases chave: función de regresión de argumento ordinal, descritores dunha ordinal

1. INTRODUCIÓN

O asunto dos habitantes centenarios (HH.CC.) é interesante para diferentes materias: bioloxía, medicina, socioloxía e economía. A duración da vida sempre ten estado presente na demografía como ciencia intersección das outras catro. O seu interese na estatística pública está relacionado coa súa influencia na función de supervivencia e o cálculo de indicadores demográficos, nomeadamente, os cocientes de dependencia de idosos H≥65/H[16,65) e de sobreavellentamento H≥85/H≥65 e mais as esperanzas de vida a diferentes idades. Tanto esa función como estas esperanzas repercuten nas seguranzas pública e privada, v.g., as rendas perpetuas e as prestacións sanitarias.

2. FONTES E MÉTODOS

A fonte de información principal que se emprega neste traballo procede do Padrón Municipal de Habitantes (PMH) do ano 2016. O PMH é o rexistro administrativo onde figuran os habitantes que teñen residencia en Galicia. No ano 2017, a cada un dos habitantes deste rexistro asignóuselle unhas coordenadas xeográficas seguindo a metodoloxía [1] e esta xeorrefenciación é a que se emprega neste relatorio para localizar os HH.CC. nas celas de 1Km^2.

Habemos denotar con q un cadro (‘quadrat’, ‘quadro’) ou cela do conxunto de cadros Q, con $C(q)$ o número de HH.CC. e $H(q)$ o número de habitantes dese cadro. O número de cadros so estudo é $\#Q = 30\,733$ mentres a superficie de Galicia é aprox. 29 575 km^2 polo que entre 1 100

e 1 200 dos cadros son incompletos. Non é un número significativo a nivel galego, menos do 5%. É rechamante (chamativo) que 11 923 deles estean baleiros.

Á partida, comprobamos se a distribución de HH.CC. nos cadros seguía unha Poisson. Nin axustando polo estimador de máxima verosemellanza nin por outros métodos, podiámola soster e houbo que rexeitar esta hipótese. Inicialmente cremos que era pola frecuencia de cero HH.CC., no entanto illando o cero tampouco seguía unha Poisson. Chegamos a conclusión de que a súa distribución é unha mestura dos cadros sen HH.CC. (aproximadamente o 95%) e dunha potencial, en concreto de orde 2. A aba superior é pesada de máis. Podemos estar perante un exemplo de azar bravo 'hasard sauvage'.

	Mulleres					
Homes	0	1	2	3	4 e máis	Total
0	17854	612	44	13	16	18539
1	175	35	9	4	11	234
2	4	4	3	4	6	21
3	0	0	0	0	8	8
4 e máis	0	0	0	1	7	8
Total	18033	651	56	22	48	18810

Tabela 1: Distribución dos cadros segundo o número de centenarias e de centenarios. 2016

No que respecta á comparanza de xéneros (tabela 1), a distribución das centenarias por cadro domina estocasticamente á dos centenarios. Existe unha relación entre as distribucións de centenarios e de centenarias como cabería esperar. As esperanzas condicionadas polos diversos valores do outro xénero son diferentes. Tamén se analizou o cociente de xéneros 'sexratio' (que para toda a poboación é de 0'2952) nos diferentes cadros considerando que se deberían estudar os menores de 0'20 e maiores de 0'40.

Conxecturamos que aba podía estar a indicar unha distribución de HH.CC. residentes en aloxamentos colectivos ou ben que o seu número dependía dalgún xeito dunha variábel de tamaño tipo poboación. Para tratar isto segundo, construímos unha función, que podemos denominar tabulamin, que calculaba o mínimo da poboación dos cadros segundo o seu número de HH.CC, $T_{min}(i) = \min_{\{q \in Q : C(q)=i\}} H(q)$. Se os HH.CC. estivesen distribuídas ao azar polo territorio galego, esperaríamos que $T_{min}(\cdot)$ fose monótona aproximadamente. Pode que estivese influída por atípicos e conviñese considerar unha medida de localización máis resistente, mais non foi preciso, xa mostra algunha anomalía ao redor do 10.

Algúns deses cadros interrompían claramente a monotonía. Foi investigada a distribución espacial dos HH.CC. polo cadro, dela resultou que non sempre habitaban nun aloxamento colectivo. Mesmo dous dos cadros correspondían a un mesmo concello de menos de 10 000 hab., polo que tras esta proba, realizouse unha comparanza da distribución das idades dos HH.CC destes cadros coa de toda Galicia. Chegando a que non era significativa ao 5% mais por pouco, $p = 0'0689$. Se ademais comparabamos a distribución dos dous cadros cos de Galicia sen eles achegábase máis a rexión de rexeitamento do estatístico χ^2, $p = 0'0603$. Nestes dous

cadros non só observamos un número anómalo de HH.CC. senón que ademais algúns deles convivían con outro centenario nun aloxamento non-colectivo.

Avanzamos na exploración cunha clasificación conxunta dos cadros por número de HH.CC. e por unha orde apriorística de urbanización empregando os conglomerados do grao e subgrao de urbanización utilizados en [2] e engadindo unha categoría que incluía os cadros non incluídos na mostra do censo de 2011 realizada polo INE (denominada Non-censo na tabela 2). Calculamos a función de regresión de argumento ordinal e observamos un bo comportamento de non-censo mais unha rotura da monotonía desta aplicación. Resolvemos isto colapsando as dúas clases (conglomerado semiurbano de 1ª e de 2ª categoría) para dispor dunha nova orde empírica que respecta a monotonía tanto da función de regresión como da de varianza, obsérvese as últimas columnas da tabela 2. Aínda que existe relación, a ordinal empírica (Urbanización nesta tabela) só explica a cuarta parte da varianza do número de HH.CC. por cadro.

Esta nova clasificación, da que só se ofrece un resumo na tabela 2, permitiu detectarmos máis candidatos a 'outliers'. Estes cadros caracterizábanse por dispor de aloxamentos colectivos neles ou nos cadros veciños. Aínda que nalgún deles, os HH.CC. residían nestes aloxamentos, na maioría non. Os que tiñan maior número de HH.CC. correspondían ás cidades e dos que estaban ao redor do 10, algúns correspondían a capitais de comarca ('cluster' de concellos).

Urbanización	0	1	2	3	4-8	9 e máis	Total	E(·|Urb.=)	D^2(·|Urb.=)
Hd	32	32	8	9	31	12	124	3,532	23,862
Urbano	519	116	35	11	9	4	694	0,434	1,211
Semi-urbano	1156	102	8	3	2	0	1271	0,108	0,166
Rural	7691	364	22	3	2	0	8082	0,053	0,062
Non-censo	8456	173	10	0	0	0	8639	0,022	0,024
Total	17854	787	83	26	44	16	18810	0,080	0,336
I.P. de Silva	0,164	0,332	0,527	0,750	0,892	0,938			

Tabela 2: Distribución dos cadros por número de HH.CC. e escalón de urbanización 2016

Pode admitirse que a distribución dos cadros de non-censo segue unha Poisson. A das celas rurais case segue unha Poisson, depende da agregación empregada no contraste khi-cadrado. Neste senso, compárese tamén, escalón a escalón, as dúas derradeiras.

Na figura 1 representamos a distribución dos cadros segundo as anteditas variábeis discreta e ordinal. Obsérvese a clara monotonía aproximada no caso de alta densidade (Hd) ata o 8 e que non se presenta columna para determinados valores na aba a partir do 11.

Podemos empregar como descritor as proxeccións sobre o eixo de ordenadas de cada columna, isto é, as frecuencias acumuladas recuando –a suma sucesiva en orde inversa $(f_C, f_C + f_{C-1}, \cdots)$– agás a derradeira que sempre é 1, entón redundante. No noso caso, a Urbanización ten $C = 5$ polo tanto ha ser unha secuencia de 4 componentes. Observamos que se calculamos a media desta serie estatística obtemos o indicador de posición de Silva (1997), IP_S, por exemplo, para $i = 1$ ($\searrow F_c(1): c = 5, \cdots, 2) = (0'04, 0'19, 0'32, 0'78)$ e o $IP_S = 0'332$.

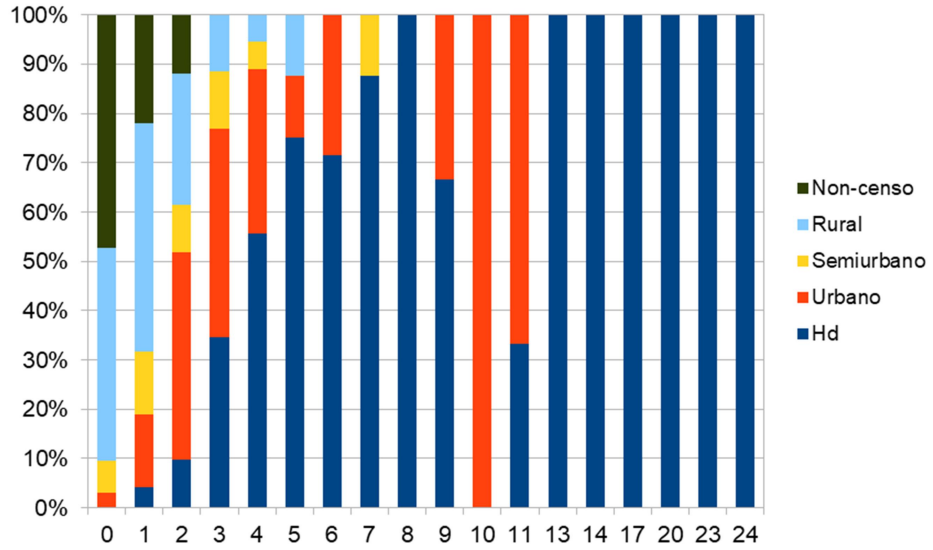


Figura 1: Distribución dos cadros por Urbanización segundo nº HH.CC.

3. CONCLUSIÓNS

Obsérvese que a función tabulamin, que resultou tan ilustrativa, había ter sentido mesmo cun carácter ordinal aínda que as conclusións tiradas dela habían ser máis febles. Os perfís que ofrece a figura 1, sobre todo o correspondente a Hd, confirman o detectado á partida pola tabulamin e enriquecen o indicador de posición que os resume.

Estamos en presenza do fenómeno de que cunha "mostra" grande (poboación finita de cadros) case todo é significativo e cunha pequena case nada é significativo [3] de aí o recurso a outras técnicas de tratamento non-paramétrico, ou sen distribución 'distribution-free', elemental mais eficaz para a validación destes datos.

Antes dunha análise espacial(-temporal) máis profunda, convirían a macrodepuración, i.e., o estudo das anomalías detectadas e a súa confirmación ou depuración, se callar, nomeadamente nalgunhas capitais de comarca porque o azar 'sauvage' podería deberse parcialmente á man do home, á inercia administrativa.

Referencias

[1] López Vizcaíno, M.E. e Iglesias Patiño, C. L. (2017). «Estratexia de xeorreferenciación da poboación de Galicia empregando técnicas estatísticas» *XIII Congreso Galego de Estatística e Investigación de Operacións*, Ferrol.

[2] IGE Grao de urbanización 2016 (GU 2016) http://www.ige.eu/estatico/pdfs/s3/clasificacions/urbanizacion/MetodoloxiaGU2016eAnexo_Modificada.pdf

[3] Silva Ayçaguer, L.C. (1997). *Cultura estadística e investigación científica en el campo de la salud. Una mirada crítica*. Editorial Díaz de Santos, Madrid.

III Encuentro Luso-Galaico de Biometría

Aveiro, 28 - 30 de Junio de 2018

DISTRIBUCIÓN ESPACIO-TEMPORAL DE LA MORTALIDAD POR INFARTO AGUDO DE MIOCARDIO EN GALICIA

María José Ginzo Villamayor[1], María Isolina Santiago Pérez[2], María Esther López Vizcaíno[3] y Rosa Mª Crujeiras Casais[1]

[1] Universidade de Santiago de Compostela, mariajose.ginzo@usc.es, rosa.crujeiras@usc.es
[2] Dirección Xeral de Saúde Pública, soly.santiago.perez@sergas.es
[3] Instituto Galego de Estatística, esther.lopez@ige.eu

RESUMEN

En este trabajo se analiza el patrón espacial y espacio temporal, por municipios, del riesgo de mortalidad por infarto agudo de miocardio (IAM) en Galicia entre los 35 y los 84 años durante el período 2000-2015.

Palabras y frases clave: Besag-York-Mollié, INLA, Infarto agudo de miocardio, mortalidad.

1. INTRODUCCIÓN

La Organización Mundial de la Salud (OMS) considera que las enfermedades cardiovasculares son la primera causa de defunción en el mundo. En el año 2012, 17,5 millones de personas murieron por enfermedades cardiovasculares. En Galicia, entre el año 2000 y 2015, se produjeron 16412 muertes por infarto agudo de miocardio (IAM).

El IAM tiene origen multifactorial. Los factores de riesgo clásicos están relacionados con estilos de vida y, por tanto, son modificables: la dislipemia, el tabaquismo, la hipertensión arterial y la diabetes. Además, en los últimos años se ha relacionado el IAM con la obesidad y el consumo de alcohol. La distribución de estos factores puede presentar diferencias geográficas en Galicia, por lo que el riesgo de sufrir un IAM, así como la mortalidad por esta causa, también pueden variar de unas zonas a otras.

El objetivo de este trabajo es analizar el patrón espacial y espacio temporal, por municipios, del riesgo de mortalidad por infarto agudo de miocardio (IAM) en Galicia entre los 35 y los 84 años durante el período 2000-2015.

2. MATERIAL Y MÉTODOS

Se analizaron las defunciones por IAM ocurridas entre los 35 y los 84 años en la población residente en Galicia en el período 2000-2015. El IAM corresponde al código I21 de la 10ª Clasificación Internacional de enfermedades (CIE-10). Los datos de mortalidad proceden del

Registro de Mortalidad de Galicia, y las poblaciones se obtuvieron del Padrón Municipal de Habitantes.

Para cada año del período de estudio se calcularon tasas específicas por edad (35 a 64 años, 65 a 74 años, 75 a 84 años), así como tasas brutas y ajustadas por edad por el método directo usando como población estándar el Censo de Galicia de 2011. Las tasas se expresan por 100.000 habitantes.

Para el análisis espacial y espacio-temporal se consideraron como unidades de análisis los municipios de Galicia, que hasta el año 2013 eran 315. En ese año se produjo la fusión de Oza y Cesuras, por lo que las defunciones ocurridas en el nuevo municipio entre 2013 y 2015 se repartieron entre los dos municipios fusionados, de forma proporcional a su población, para mantener los 315 municipios en todo el período de estudio.

Fijando como regiones administrativas los municipios en este trabajo se usan métodos espaciales y espacio-temporales para el análisis de datos de conteo que permiten modelar el patrón subyacente al número de defunciones por infarto en Galicia. Estos métodos son útiles para caracterizar patrones de muertes por infarto formulados mediante modelos jerárquicos.

Para el ajuste de los modelos jerárquicos en este contexto se utiliza la metodología Integrated Nested Laplace Approximation (INLA) propuesta por Rue et al. (2009).

Se presentan a continuación los dos modelos estudiados, espacial y espacio-temporal, empleados en este trabajo.

2.1. Modelización espacial

Para analizar el patrón espacial y espacio temporal de las muertes por infarto en Galicia, considerando los efectos de distintas covariables, se ajustó el modelo propuesto por Besag et al. (1991), adaptado a este contexto. Uno de los supuestos de este modelo es que el log-riesgo se puede descomponer como suma de una componente espacial estructurada y un error aleatorio, pero también se puede incluir el efecto suave de alguna covariable.

Se denota por $\mathrm{Y_i} =$ el número de personas que fallecieron por IAM en el municipio i, para cada $\mathrm{i} = 1, \ldots, \mathrm{n}$

Este proceso de conteo será modelado a través de un modelo Poisson-LogNormal (ver Banerjee et al. (2004)). Es decir

$$\mathrm{Y_i} \mid \eta_\mathrm{i} \sim \mathrm{Pois}\left(\mathrm{E_i} \exp\left(\eta_i\right)\right)$$

donde E_i es la población en riesgo, η_i (los riesgos log-relativos) un predictor lineal y las variables $\mathrm{Y_i}$ condicionalmente en η_i son independientes. El campo latente η_i s tendrá en cuenta para modelar la estructura subyacente y recoger la variabilidad espacial. Una formulación sencilla viene dada por:

$$\eta_i = \mu + f_v(s_i) + f_u(s_i)$$

donde s_i es el centroide de cada municipio y f_v y f_u denotan los efectos espaciales estructurado y o no estructurado, respectivamente. Para f_s, se impondrá un campo aleatorio Gaussiano de Markov (GMRF) intrínseco (Rue y Held (2005)). Para f_u, se considera el proceso de ruido blanco que representa la "sobredispersión" que pueden presentar los municipios:

- Denotando por $z(s_i) \equiv f_v(s_i), i = 1,....,n$, Z es un GMRF. Los $z(s_i), z(s_j)$ con $i \neq j$ si son dependientes con estructura de Markov y siguen una distribución $N\left(0, \iota_2\right)$.
- Por otro lado, se denota por $w(s_i) \equiv f_u(s_i), i = 1,....,n$, W es ruido blanco. Los $w(s_i), i = 1,....n$ son independientes con distribución $N\left(0, \iota_1\right)$.

2.2. Modelización espacio-temporal

Investigando solo el patrón espacial de los IAM no nos permite concluir nada sobre otra de las componentes de variación, la temporal, que puede ser igualmente de interés. El modelo anterior pode extenderse fácilmente al caso espacio-temporal incluyendo el tiempo. La variable respuesta para un municipio i será:

Y_i = el número de personas que fallecieron por IAM en el municipio i en el tiempo t

que será observada en los n municipios y en T instantes de tiempo. El modelo espacial anterior se extiende al permitir la componente temporal quedando:

$$Y_{it} \mid \eta_i \sim \text{Pois}(E_{it} \exp(\eta_{it}))$$

A formulación que seguirá, en este caso, el campo latente es:

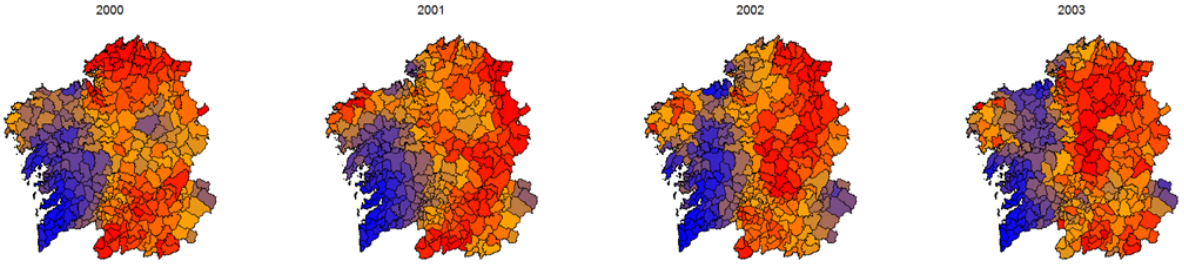
$$\eta_{it} = \mu + f_v(s_i) + f_u(s_i) + f_T(t)$$

Con $t = 1,...,T$, donde en $f_T(t)$ se especifica la estructura temporal. Dicha estructura puede corresponder a una componente lineal en el tiempo, a una componente suave o bien a un proceso que incluya correlación temporal.

Como resultados, se presentan los mapas de las componentes estructuradas y se realiza un análisis del efecto de las covariables.

3. RESULTADOS

En el período 2000-2015 se produjeron en Galicia 16.412 defunciones por IAM entre la población de 35 a 84 años, lo que supone un 6% de todas las defunciones ocurridas en ese período y en ese grupo de edad. Un 68% de las muertes (11.204) ocurrieron en hombres, y un 51% (8.335) en el grupo de edad de 75 a 84 años. En todos los años del período de estudio y en todos los grupos de edad se observa mayor mortalidad en los hombres frente a las mujeres, y el número de muertes aumenta con la edad. La mortalidad por IAM entre los 35 y los 85 años descendió en Galicia en el período 2000 a 2015, con una tasa bruta de 88,5 defunciones por 100.000 hab. en el año 2000 y de 43,5 por 100.000 en el año 2015. El descenso se produjo paulatinamente durante todo el período de estudio y fue más acusado en los hombres. La mortalidad por IAM también descendió en todas las edades durante el período de estudio, con un cambio relativo anual del 6,7% en el grupo de 35 a 64 años, de 7,6% en el de 65 a 74 y de 5,7% en el de 75 a 84 años. En este grupo la tasa de mortalidad pasó de 352,5 por 100.000 en el año 2001 a 159,8 por 100.000 en el año 2015, pero también en los otros dos grupos la tasa se redujo a la mitad en estos 15 años. El patrón espacial del riesgo de mortalidad por IAM se presenta en la Figura 1. En los años analizados, el norte de la provincia de Coruña, la provincia de Lugo y Ourense son las que tienen un mayor riesgo de mortalidad por infarto, aunque se observan diferencias entre los distintos años.



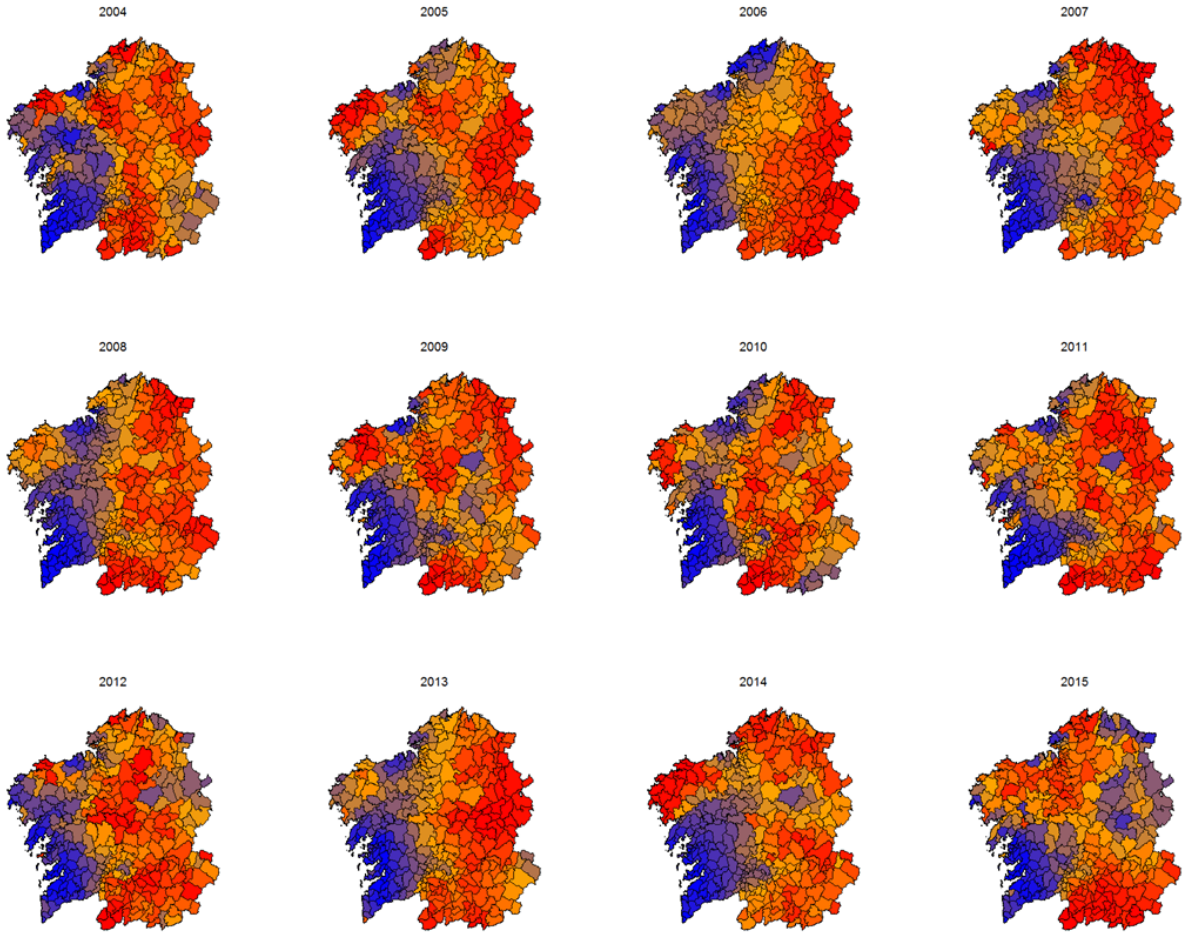


Figura 1. Distribución geográfica del número de infartos. En color rojo se indican las zonas de mayor riesgo y en azul las zonas de menor. De izquierda a derecha y de arriba abajo se tienen los datos para cada año. Así por ejemplo el año de la posición (2,2) es el 2005.

AGRADECIMIENTOS

María J. Ginzo y Rosa M. Crujeiras agradecen el apoyo del proyecto MTM2016-76969-P del Ministerio de Economía y Competitividad.

Referencias

[1] Banerjee S., Carlin, B. E., Gelfand, A. (2004) Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC Monographs on Statistics & Applied Probability

[2] Besag, J., York, J. e Mollié, A. (1991) Bayesian image restoration with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics, 43, 1-59.

[3] Rue, H., e Held, L. (2005) Gaussian Markov random fields: theory and applications. Chapman and Hall, CRC Press, London.

[4] Rue, H., Martino, S. e Chopin, N (2009) Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. Journal of the Royal Statistical Society, 71, 319-392.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

INCORPORATING SURVEY WEIGHTS IN SPATIAL MODELLING OF OBESITY AND HYPERTENSION IN SOUTH AFRICA

Sheyla Rodrigues Cassy[1,4], Samuel O Manda[2,3], Filipe Marques[4], M Rosário Martins[5] and Pedro Silva[6]
sheylaratan@hotmail.com, samuel.manda@mrc.ac.za, fjm@fct.unl.pt, mrfom@ihmt.unl.pt, pedro-luis.silva@ibge.gov.br

[1]DMI, Faculdade de Ciências, Universidade Eduardo Mondlane, Maputo, Mozambique
[2]Biostatistics Research Unit, South African Medical Research Council, Pretoria, South Africa
[3]Division of Epidemiology and Biostatistics, School of Public Health, University of Witwatersrand, Johannesburg, South Africa
[4]CMA, DM, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
[5]Global Health and Tropical Medicine, GHTM, Instituto de Higiene e Medicina Tropical, IHMT, Universidade Nova de Lisboa, Lisboa, Portugal
[6]Escola Nacional de Ciências Estatísticas (ENCE), Rio de Janeiro, Brazil

ABSTRACT

Several studies in the Sub-Saharan Africa (SSA) have used complex nationally representative household and population surveys to provide hierarchically and spatially smoothed estimation and prediction of health outcomes at subnational levels. These surveys often employ a complex design that includes two-stage clustering sample and stratification, with disproportionate sampling. However, while a number of studies [2,3,6] have accounted for clustering, the sample weights are rarely considered when spatial modelling using the survey data is undertaken. We develop techniques to integrate the sampling weights for multiple health outcomes arising from complex surveys using recently developed hierarchical Bayesian multivariate spatial models [1,4]. A simulation study is conducted to demonstrate the performance of the proposed method. Using the data from the 2014-15 National Income Dynamics Study (NIDS) [5], we also analyze the prevalence of Obesity and Hypertension at the health district level in South Africa. The models were fitted within the `R` environment using the readily available packages.

Keywords and key sentences: NIDS, NCDs, complex survey design, design weights, spatial modelling.

ACKNOWLEDGMENT

This research was partially funded by the Fundação para a Ciência e a Tecnologia (Portuguse Foundation for Science and Technology) through the project UID/MAT/00297/2013 (Centro de Matemática e Aplicações). Research of Sheyla Rodrigues Cassy is funded by the Calouste Gulbenkian Foundation grant process 135422. This work of Samuel Manda was supported by the South Africa Medical Research Council (SAMRC) with funds from National Treasury in terms of the SAMRC's competitive Intramural Research Fund: SAMRC-RFA-IFF-02-2016. The funding body did not have any role in the design of the study and collection, analysis, and interpretation of data in writing the manuscript.

References

[1] Chen, C., Wakefield, J., & Lumely, T. (2014). The use of sampling weights in Bayesian hierarchical models for small area estimation. *Spatial and spatio-temporal epidemiology*, 11, 33-43.

[2] Kandala, N. B., Manda, S. O., Tigbe, W. W., Mwambi, H., & Stranges, S. (2014). Geographic distribution of cardiovascular comorbidities in South Africa: a national cross-sectional analysis. *Journal of Applied Statistics*, 41(6), 1203-1216.

[3] Manda, S., Masenyetse, L., Cai, B., & Meyer, R. (2015). Mapping HIV prevalence using population and antenatal sentinel-based HIV surveys: a multi-stage approach. *Population health metrics*, 13(1), 22.

[4] Mercer, L., Wakefield, J., Chen, C., & Lumley, T. (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial statistics*, 8, 69-85.

[5] Southern Africa Labour and Development Research Unit. *National Income Dynamics Study 2014 - 2015, Wave 4* [dataset]. Version 1.1. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2016. Cape Town: DataFirst [distributor], 2016. Pretoria: Department of Planning Monitoring and Evaluation [commissioner], 2014.

[6] Weimann, A., Dai, D., & Oni, T. (2016). A cross-sectional and spatial analysis of the prevalence of multimorbidity and its association with socioeconomic disadvantage in South Africa: a comparison between 2008 and 2012. *Social Science & Medicine*, 163, 144-156.

Comunicações em Póster

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

KIDNEY INSUFFICIENCY: A STATISTICAL ANALYSIS BASED ON THE GAMLSS FRAMEWORK

Ana Julia Righetto[1],Thiago Gentil Ramires[2], Luiz Ricardo Nakamura[3], Edwin M. M. Ortega[4] e Gauss M. Cordeiro [5]

[1]Instituto Agronômico do Paraná, Londrina-PR, Brazil
[2]Universidade Tecnológica Federal do Paraná - Apucarana-PR, Brazil
[3]Universidade Federal de Santa Catarina, Florianópolis-SC, Brazil
[4]Universidade de São Paulo, Piracicaba-SP, Brazil
[5]Universidade Federal de Pernambuco, Recife-PE, Brazil.

ABSTRACT

Renal disease is a medical and public health problem worldwide. In this work, we analysi a data set of patients with renal disease in a Brazilian city and use the generalized additive models for location, scale and shape (GAMLSS) framework based on the Weibull distribution to identify the associated factors in these patients.

Keywords: GAMLSS; Heteroscedastic models; Renal disease; Weibull.

1. INTRODUCTION

According to [1], hypertension and diabetes are the main risk factors for chronic kidney disease (CKD) and are becoming more frequent in the general population, contributing for the increased incidence of CKD. In addition to age, hypertension and diabetes, other factor of complication in patients undergoing hemodialysis are the infection of hepatitis B (HBS) and C (HCV)[3];[2]
In Brazil there are insufficient studies that evaluate, at a national level, the survival and associated factors of patients in dialysis modalities [5]. Thus, regional and local studies are essential to identify such factors and to evaluate the survival of patients with CKD. With this propose, a survey was conducted at the Kidney Institute of Maringá to predict survival and identify associated factors in the lifetime of patients with renal insufficiency from the metropolitan region of Maringá-Paraná, Brazil.

2. MATERIAL AND METHODS

2.1 Data set description

The data set has a total of 177 patients, of whom 23 underwent kidney transplant and 22 were diabetic. Most patients treated by the institute were from the Brazilian public health system and all patients in the survey were undergoing hemodialysis. The total of failure times

(death) is 119, and 58 observations were considered censored (32% of censure) in case that the patient did not continue in the program for any reason or that the patients did not die until the end of study.
The variables considered in this survey were: x_1) t_i: observed time (in days); x_2) δ_i: failure indicator (censored or observed); x_3) age (in years) at the beginning of treatment; x_4) gender[male=0;female=1]; x_5) marital status (MS); x_6) skin color indicator (SCI); x_7) : kidney transplant indicator (KTI); x_8) antibodies indicator for hepatitis C (HepC); x_9) antibodies indicator for hepatitis B (HepB); x_10) diabetic indicator (DI).

2.2 GAMLSS framework

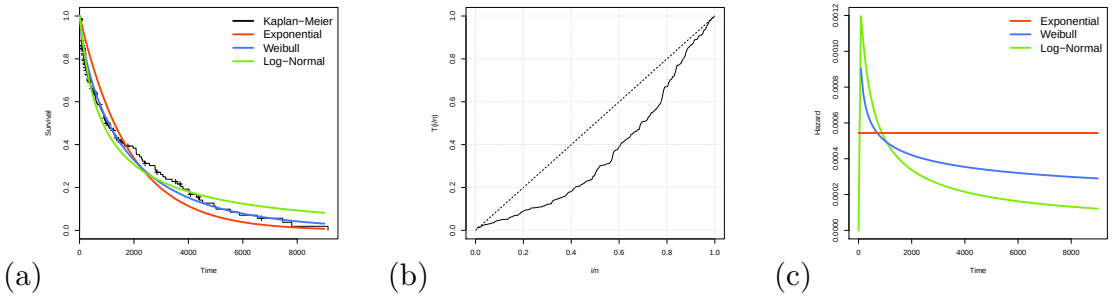
[4] proposed the generalized additive models for location, scale and shape (GAMLSS), where the systematic part of the model is expanded allowing not only the location but all the parameters of the conditional distribution of the response variable to be modelled as parametric functions of a set of explanatory variables. In this work we are only considering the parametric GAMLSS with up to two parameters, i.e., for $T \sim \mathcal{D}(t;\boldsymbol{\theta})$, where $\mathcal{D}$ represents the distribution of T and $\boldsymbol{\theta}^\top = (\mu, \sigma)$ denotes the vector of its parameters, the parametric GAMLSS model can be written as

$$g(\mu) = \boldsymbol{X}_1\boldsymbol{\beta}_1 \quad \text{and} \quad g(\sigma) = \boldsymbol{X}_2\boldsymbol{\beta}_2,$$

where $g_k(\cdot)$, $k = 1, 2$, denote appropriate link functions, $\boldsymbol{X}_k$ is a known model matrix and $\boldsymbol{\beta}_k = (\beta_{0k}, \ldots, \beta_{m_k k})^\top$ is a parameter vector. In this work we considered three different distributions as suitable candidates to explain the observed time of the patients: exponential, Weibull and log-normal. All the procedures described in this section were performed in the `gamlss` package [6] in `R`.

3. RESULTS AND DISCUSSION

We start by fitting the exponential, Weibull and log-normal models disregarding regression variables. We provide in Figure 1(a) the plots of the estimated and empirical survival functions, wherein we can conclude that the Weibull distribution provides a good fit to these data.



(a) (b) (c)

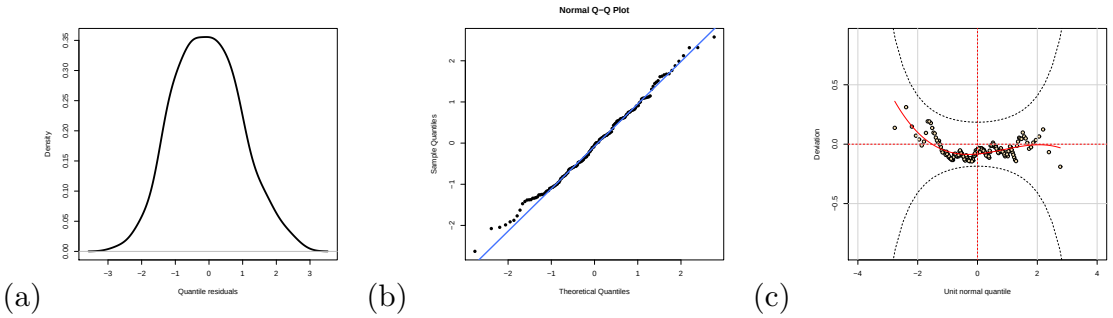
Figure 1: (a) Estimated and empirical survival functions; (b) TTT plot; and (c) estimated hazard rate function of the proposed distributions.

The TTT plot given in Figure 1(b) reveals that the hazard function has decreasing shape. On the other hand, Figure 1(c) shows that the estimated hazard functions of the exponential, Weibull and log-normal models are constant, decreasing and unimodal (increasing and then decreasing), respectively, indicating that the Weibull distribution is appropriate to fit these data. So, after selecting the Weibull distribution as the best one, we propose a Weibull regression model. Using the stepGAIC procedure, the final best model based on the Weibull distribution is given by

$$\begin{aligned}\hat{\mu} &= \exp\{8.905 - 0.036x_3 + 0.444x_4[=0] + 0.282x_7[=T] + 0.429x_8[=T] + 0.531x_9[=T]\} \qquad (1)\\ \hat{\sigma} &= \exp\{-0.795 + 0.008x_3 + 0.860x_7[=T] + 0.756x_8[=T] + 0.121x_9[=T]\}.\end{aligned}$$

From the model for the mean μ, we can note that the older is the patient when he starts treatment, the lower is its lifetime expectancy. The lifetime expectancy is lower in female patients and in patients who have not undergone kidney transplantation. Moreover, patients who have antibodies indicator to hepatitis C and B have greater lifetime expectancy.
Regarding the shape parameter σ, the variable age at the beginning of treatment has a positive linear relationship with the variability of the lifetimes of the patients, i.e. the variability of the lifetimes increases when the patients drawled to start treatment. Furthermore, the variability also increases when the patient has undergone kidney transplantation and when the presence of antibodies to hepatitis B and/or C are positive. Here, it is clear that the variability of the lifetimes of patients must be modelled as well, and not only the mean (as other more traditional models, such as classical linear regression or generalized linear model usually do), which could mislead us on the selection of false risk factors for kidney failure.
Figure 2 shows some residual plots that can help to verify the adequacy and the assumptions of the chosen fitted model. Panel (a), suggests that the normalized quantile residuals have an approximately normal distribution. Panel (b) shows that there are few points off the line in the low end of the range. Finally, on Panel (c), the worm-plot (WP) also suggests that there is a slightly problem in the lower tail of the distribution of T. The correct would be the fitted cubic model (red curve) be a line on the X-axis. Nevertheless, the Weibull regression model based on the GAMLSS framework provides a reasonable fit to these data. Finally, using the results presented in Equation (1) we can calculate the average lifetime for each explanatory variable selected in the regression model. The averages are easily calculated using the expression

$$E(T) = \exp(8.905 - 0.036x_1 + 0.444x_2 + 0.282x_5 + 0.429x_6 + 0.531x_7).$$



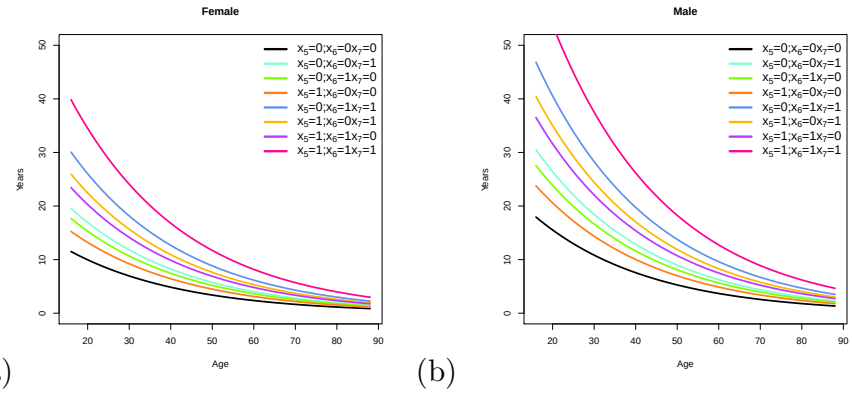
(a) (b) (c)

Figure 2: Residual from the Weibull regression model based on the GAMLSS framework: (a) Density of the quantile residuals; and (b)Q-Q plot and (c) WP.

Consider the high-risk group ($x_1 = \max(x_1) = 88$, $x_2 = 0$, $x_5 = 0$, $x_6 = 0$ and $x_7 = 0$) and the low-risk group ($x_1 = \min(x_1) = 17$, $x_2 = 1$, $x_5 = 1$, $x_6 = 1$ and $x_7 = 1$). The average of the lifetime for the high-risk and low-risk groups are 310 and 21,568 days, respectively. In Figure 3 we present the average lifetime (in year) as function of x_2 by changing the levels of x_5, x_6 and x_7 variables.

4. CONCLUSION

In the current survey, we verify based on the fit of the Weibull model that patients from metropolitan area of Maringá (Brazil), that the late onset of renal insufficiency, female gender, absence of kidney transplant and negative antibodies to hepatitis B and C can be identified as negative factors more important for the lifetime of patients. Hence, patients having such



(a) (b)

Figure 3: Expected lifetime as function of age at the beginning of treatment, for all levels of x_5, x_6 and x_7 for: (a) female ($x_2 = 0$) and (b) male ($x_2 = 1$).

characteristics require special attention. Furthermore, the age at the beginning of treatment, kidney transplant indicator and antibodies to hepatitis B and C are significant factors to explain the variability of the survival time.

References

[1] Bommer, J. (2002). Prevalence and socio-economic aspects of chronic kidney disease. *Nephrology Dialysis Transplantation*, **17**, 8-12.

[2] Ferreira, R.C.; Teles, S.A.; Dias, M.A.; Tavares, V.R.; Silva, S.A.; Gomes, S.A.; Yoshida, C.F.T.; Martins, R.M.B. (2006). Hepatitis B virus infection profile in hemodialysis patients in Central Brazil: prevalence, risk factor, and genotypes. *Memórias do Instituto Oswaldo Cruz*, **101**, 689-692.

[3] Leão, J.R.; Pace, F.H.D.L.; Chebli, J.M.F. (2010). Infecção pelo vírus da hepatite C em pacientes em hemodiálise: prevalência e fatores de risco. *Arquivos de Gastroenterologia* **47**, 28-34.

[4] Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* **54**, 507–554.

[5] Sancho, L.G.; Dain, S. (2008). Análise de custo-efetividade em relação as terapias renais substitutivas: como pensar estudos em relação a essas intervenções no Brasil. *Cadernos de Saúde Pública*, **24**, 1279-1290.

[6] Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software,* **23**, 1–46.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

INFEÇÕES POR PROTOZOÁRIOS INTESTINAIS E DÉFICE DE CRESCIMENTO EM LACTENTES DE SÃO TOMÉ: UM ESTUDO DE COORTE DE NASCIMENTO

Marta Alves[1,2], Ana Luisa Papoila[1,2,3], Marisol Garzón[4] e Luís Pereira-da-Silva[1,5]

[1]Centro de Investigação do Centro Hospitalar de Lisboa Central
[2]CEAUL - Centro de Estatística e Aplicações da Universidade de Lisboa
[3]NOVA Medical School/Faculdade de Ciências Médicas da Universidade NOVA de Lisboa
[4]Tropical Clinic Teaching and Research Unit, Instituto de Higiene e Medicina Tropical, Universidade NOVA de Lisboa
[5]Medicine of Woman, Childhood and Adolescence Teaching and Research Area, NOVA Medical School/Faculdade de Ciências Médicas da Universidade NOVA de Lisboa

RESUMO

Em lactentes de países de baixo e médio rendimento, *Giardia lamblia*, *Cryptosporidium* spp. e helmintas são agentes prevalentes em infeções intestinais. As interações hospedeiro-parasita podem levar a uma resposta inflamatória da mucosa e aumento da permeabilidade intestinal. Clinicamente, isto pode refletir-se por impacto negativo no crescimento e neurodesenvolvimento. Os efeitos destas infeções intestinais subclínicas na saúde infantil têm sido pouco estudados. Este projeto teve como objetivo analisar as associações entre infeções por parasitas intestinais em crianças assintomáticas de São Tomé e o seu estado nutricional e, ainda, obter estimativas das curvas de crescimento de referência no que diz respeito ao comprimento-para-idade e ao peso-para-idade. Para este efeito, foram utilizados modelos de regressão aditivos generalizados de efeitos mistos e modelos de regressão aditivos generalizados para localização, escala e forma.

Palavras e frases chave: crescimento infantil, protozoários intestinais, modelos aditivos generalizados de efeitos mistos, modelos aditivos generalizados para localização, escala e forma.

1. INTRODUÇÃO

Os primeiros 1000 dias de vida - o período entre a concepção e o final do segundo aniversário – constituem um período único de oportunidades em que se estabelecem os alicerces para atingir condições ideais de saúde, crescimento e neurodesenvolvimento ao longo da vida (Cusick et al., 2016). Durante este período, a velocidade de crescimento linear é mais rápida do que em qualquer outra fase da vida, incluindo a adolescência. O crescimento precoce tem um papel importante na programação da trajetória de crescimento durante a infância e adolescência e na estatura na idade adulta (Matorell et al., 2017).

Nos países de baixo e médio rendimento é amplamente reconhecido que a restrição do crescimento começa na vida intrauterina e continua pelo menos nos dois primeiros anos de vida. Nestes países, o défice de crescimento fetal ou do crescimento linear nos dois primeiros anos de vida levam a consequências irreversíveis, incluindo menor estatura na idade adulta, menor escolaridade e redução do peso ao nascer dos descendentes (Victora et al., 2010).
Neste contexto, o estudo multinacional MAL-ED (Malnutrition and Enteric Disease) realizado em oito países de baixo e médio rendimento destacou o papel dos protozoários como agentes etiológicos das infeções entéricas nos dois primeiros anos de vida (Platts-Mills et al., 2015). Vários outros estudos de coorte de nascimento desenvolvidos em países de baixo e médio rendimento, providenciaram evidência de que a *Giardia lamblia* (Donowitz et al., 2016), a *Cryptosporidium* spp. (Korpe et al., 2016) e as infeções por helmintas (Gyorkos et al., 2011) durante os primeiros dois anos de vida têm impacto no crescimento.

Este projeto teve como objetivo estudar a associação entre infeções por parasitas intestinais e o crescimento de crianças assintomáticas de São Tomé, do nascimento aos 24 meses de idade. Adicionalmente, foram estimadas curvas de crescimento de referência para o comprimento-para-idade e para o peso-para-idade, para esta população.

2. MÉTODOS

Este é um estudo de coorte de nascimento realizado na ilha de São Tomé pertencente ao arquipélago de São Tomé e Príncipe, entre março e junho de 2013, com seguimento até aos 24 meses de idade. Os detalhes desta coorte de nascimento são descritos em Garzón et al. (2017). Ressalva-se que os recém-nascidos foram recrutados no período neonatal, ou seja, durante os primeiros 28 dias pós-natais.
Foram elegíveis os recém-nascidos adequados para a idade gestacional (percentis$>$ 10 e $<$90). Foram excluídos os recém-nascidos com baixo peso ($<$2500 g), prematuros ($<$37 semanas de gestação), sem informação sobre a idade gestacional, com malformações congénitas graves ou com necessidade de hospitalização por asfixia perinatal.
Foram consideradas variáveis epidemiológicas, socioeconómicas, de práticas alimentares e parâmetros clínicos. O índice de pobreza multidimensional (IPM), que inclui as dimensões da educação, saúde e padrão de vida (Alkire et al., 2010), foi usado para avaliar a condição socioeconómica do agregado familiar. A antropometria foi avaliada mensalmente e incluiu o comprimento-para-idade, o peso-para-idade, o crescimento atingido em *z-scores* (para peso/comprimento - WLZ, comprimento/idade – LAZ), a diferença do comprimento-para-idade (LAD), a velocidade de crescimento ponderal (peso) e linear (comprimento) em *z-scores* (WAVZ e LAVZ) e a desnutrição (aguda e crónica, definida como $<$ 1 Desvio Padrão). A presença de protozoários e helmintas intestinais foi avaliada trimestralmente por técnicas microscópicas.

Para explorar a associação entre cada parâmetro antropométrico e as infeções parasitárias entéricas e outras covariáveis relevantes (e.g. IPM, estatura da mãe, práticas alimentares e eventos clínicos agudos) foram utilizados modelos de regressão aditivos generalizados de efeitos mistos que têm em consideração a estrutura de autocorrelação entre as medidas ao longo do tempo. No estudo multivariável, a idade foi modelada com *splines* devido à sua associação não-linear com cada parâmetro antropométrico.
Para obter as estimativas das curvas de crescimento foram aplicados modelos aditivos generalizados para localização, escala e forma, de acordo com a abordagem seguida pela Organização Mundial de Saúde (WHO Child Growth Standards, 2016). Na modelação, foram utilizadas as distribuições *Box-Cox power exponential*, *Box-Cox t*, *Box-Cox normal* e *Johnson's SU* e, no que diz respeito à suavização, foram aplicados *splines* cúbicos e polinómios fracionários. O melhor modelo foi escolhido com base no critério de informação de Akaike generalizado,

nos *worm plots* e no *Q-test* (Stasinopoulos et al., 2017).
O protocolo do estudo foi aprovado pela comissão de ética do Ministério da Saúde da República Democrática de São Tomé e Príncipe, e foi obtido o consentimento informado dos tutores legais de todas as crianças.

3. RESULTADOS

Foram incluídos 475 recém-nascidos, representando 16,5% dos nados-vivos em São Tomé tendo 280 (58,9%) completado os 24 meses de seguimento. *Giardia lamblia* e helmintas foram os parasitas mais prevalentes. A análise multivariável revelou que a infeção por *Giardia lamblia* e helmintas se associou de uma forma significativa com a diminuição do crescimento linear (LAZ : $\hat{\beta}$=-0,10 IC 95%: -0,18 a -0,02 e $\hat{\beta}$=-0,16 IC 95%: -0,25 a -0,07, respetivamente e LAD: $\hat{\beta}$=-0,32 IC 95%: -0,57 a -0,07 e $\hat{\beta}$=-0,48 IC 95%: -0,76 a -0,20, respetivamente). A infeção por *Cryptosporidium* spp. associou-se significativamente a uma diminuição na velocidade de crescimento ponderal (WAVZ: $\hat{\beta}$=-0,43 IC 95%: -0,80 a -0,06) e linear (LAVZ: $\hat{\beta}$=-0,55 IC 95%: -0,94 a -0,17).

4. CONCLUSÕES

Este é o primeiro estudo de coorte de nascimento em São Tomé, pioneiro em estudar associações entre o crescimento e as infeções por parasitas intestinais. As infeções, inclusive subclínicas, revelaram estar associadas significativamente à restrição do crescimento. Estas associações são problemáticas em São Tomé, endémico para *Giardia lamblia* e helmintas, em contexto de proporção não negligenciável de lactentes marginalmente desnutridos. Estes poderão ter capacidade limitada para reparar lesões da mucosa, com impacto negativo no crescimento.

AGRADECIMENTOS

Este projeto foi financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e Tecnologia - no âmbito dos projetos SFRH/BD/81431/2011 e UID/MAT/00006/2013. Os autores agradecem ainda o apoio prestado pela organização não governamental Instituto Marquês de Valle Flôr.

Referências

[1] Alkire, S., Santos, M.E. (2010). *Acute Multidimensional Poverty: A New Index for Developing Countries*. Working Papers 38; OPHI Oxford Poverty & Human Development Initiative, University of Oxford: Oxford, UK.

[2] Cusick S.E., Georgieff M.K. (2016). The Role of Nutrition in Brain Development: The Golden Opportunity of the "First 1000 Days". *J Pediatr.* 175, 16–21.

[3] Donowitz J.R., Alam M., Kabir M., Ma J.Z., Nazib F., Platts-Mills J.A., Bartelt L.A., Haque R., Petri W.A. (2016). A Prospective longitudinal cohort to investigate the effects of early life giardiasis on growth and all cause diarrhea. *Clin. Infect. Dis.* 63, 792–797.

[4] Garzón M., Pereira-da-Silva L., Seixas J.,; Papoila A.L., Alves M., Ferreira F., Reis A. (2017). Association of enteric parasitic infections with intestinal inflammation and permeability in asymptomatic infants of São Tomé Island. *Pathog. Glob. Health*, 111, 116–127.

[5] Gyorkos T.W., Maheu-Giroux M., Casapía M., Joseph S.A., Creed-Kanashiro H. (2011). Stunting and helminth infection in early preschool-age children in a resource-poor community in the Amazon lowlands of Peru. *Trans. R. Soc. Trop. Med. Hyg.* 105, 204–208.

[6] Korpe P.S., Haque R., Gilchrist C., Valencia C., Niu F., Lu M., Ma J.Z., Petri S.E., Reichman D., Kabir M. et al. (2016). Natural history of Cryptosporidiosis in a longitudinal study of slum-dwelling Bangladeshi children: association with severe malnutrition. *PLoS Negl. Trop. Dis.* 10, e0004564.

[7] Martorell R. (2017). Improved nutrition in the first 1000 days and adult human capital and health. *Am J Hum Biol.* 29(2).

[8] Platts-Mills J.A., Babji S., Bodhidatta L., Gratz J., Haque R., Havt A., McCormick B.J., McGrath M., Olortegui M.P., Samie A., et al. (2015). Pathogen-specific burdens of community diarrhea in developing countries: A multisite birth cohort study (MAL-ED). *Lancet Glob. Health* 3, e564–e575.

[9] Stasinopoulos, D. M., Rigby, R.A., Heller, G., Voudouris, V., De Bastiani, F. (2017). *Flexible Regression and Smoothing: Using GAMLSS in R.* Chapman and Hall/CRC.

[10] Victora C.G., de Onis M., Hallal P.C., Blössner M., Shrimpton R. (2010). Worldwide timing of growth faltering: revisiting implications for interventions. *Pediatrics.* 125, e473–80.

[11] WHO Multicentre Growth Reference Study Group (2006). *WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development.* Geneva: World Health Organization.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

FATORES QUE CONDICIONAM A ACEITAÇÃO DA DIRETIVA DA LINHA DE SAÚDE 24 DE NÃO IR A UM SERVIÇO DE URGÊNCIAS

Isabel Natário[1], Paula Simões[2], Joaquim Pina[3] e Sérgio Gomes[4]

[1]CMA; Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal; icn@fct.unl.pt
[2]CMA; Área Departamental de Matemática, ISEL - Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa,Portugal; paulasimoes@adm.isel.pt
[3]CEFAGE-FCT/UNL; Departamento de Ciências Sociais Aplicadas, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal; jagl@fct.unl.pt
[4]Direção Geral de Saúde, Portugal; sergiogomes@dgs.pt

RESUMO

A Linha de Saúde 24 (SNS24) é uma linha nacional de atendimento telefónico que apoia o cidadão em diversas questões relacionadas com a saúde. Quando a questão se relaciona com episódios em que se o cidadão não tivesse ligado teria ido procurar ajuda num serviço de urgências hospitalares e, adicionalmente, se a chamada o desviou dessa intenção inicial (porque possivelmente lhe foi indicada uma terapêutica em casa ou um outro serviço de atendimento ao médico sem o carácter de urgência), tal é benéfico para os hospitais. O serviço de urgências fica aliviado e os hospitais poupam direta e indiretamente, respetivamente por não atenderem o paciente ou por não verem os seus recursos comprometidos em alturas de pico de certas doenças sazonais, por exemplo, e evitando potenciais contágios desnecessários. As admissões nas urgências constituem um dos mais expressivos fatores dos custos hospitalares, que podem ser mitigados com recursos deste género.

É contudo importante avaliar o impacto económico deste tipo de alternativa e para tal é essencial quantificar os fatores que influenciam, por um lado, o uso da linha SNS24 e, por outro, os que contribuem para que um seu utente altere a sua decisão inicial de se dirigir a um serviço de urgência. O primeiro ponto foi já tratado nos artigos de Simões *et al.* [6, 7]. Este trabalho concentra-se no segundo aspeto.

Para os dados envolvidos neste estudo a sua localização pode ser um fator relevante assim como a sua evolução no tempo, pelo que os modelos a empregar deverão levar estes pontos em consideração. Assim, vai recorrer-se à modelação de dados binários através de modelos de regressão logística espaciais e/ou espaço-temporais, abordagem pouco frequente para dados desta natureza não-geoestatísticos. Verificando que os resíduos de um modelo de regressão logística exibem correlação espacial esta deve ser levada em conta na modelação. Para

tal, as abordagens mais comuns são regressões com coeficientes localmente estimados em modelos de regressão logística geograficamente ponderados [3, 9], o uso de covariáveis espacial e/ou temporalmente variáveis [4], o uso de modelos aditivos generalizados com efeitos não paramétricas na coordenadas geográficas da localização dos dados [1], o uso de efeitos aleatórios no contexto de modelos mistos [2, 9]. Contudo, há problemas inerentes, que poderão ser contornados pela utilização de modelos de regressão de processos Gaussianos para dados binários [8].

Neste estudo faz-se uma primeira análise dos dados binários da mudança de opinião, quanto à ida a um serviço de urgência, dos utentes com essa intenção que ligaram para a linha SNS24 entre 2010 e 2016. Para tal reveem-se as técnicas disponíveis, adaptando e propondo um conjunto de ideias para que sejam úteis, visando os objetivos propostos. Apesar dos resultados não serem transversalmente concordantes entre metodologias, o sexo, a idade, o subsistema de saúde e a morada dos utentes revelam-se fatores importantes na probabilidade de mudança de opinião em modelação. A relevância da componente espacial inerente pode indicar que não estão a ser considerados outros fatores ainda importantes, para os quais não se detém informação como, por exemplo, características intrínsecas do utente como a sua escolaridade, a sua literacia, os seus rendimentos ou outros. O tempo não é relevante na modelação, possivelmente devido a uma série temporal curta.

Palavras-chave: Linha de Saúde 24 (SNS24), regressão logística espacial, auto-correlação espacial.

AGRADECIMENTOS

Este trabalho é financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito dos projetos UID/MAT/00297/2013, UID/MAT/00006/2013 e UID/ECO/04007/2013, e por Fundos FEDER através do Programa Operacional Fatores de Competitividade – COMPETE (POCI-01-0145-FEDER-007659).

Referências

[1] Almeida, C.T., Nascimento-Jr, J.R.O., Neto, J.S., Lorenzon, M.C.A., Tassinari, W.S. (2010). Utilização de Modelos Logísticos Espaciais aplicados na análise da Sanidade Apícola do Estado do Rio de Janeiro. *In: Simpósio de Pesquisa Operacional e Logística da Marinha, 2010, Rio de Janeiro, Anais do SPOLM 2010.*

[2] Banerjee, S., Carlin. B.P., Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall, Boca Raton.

[3] Brunsdon, C., Fotheringham, A., Charlton, M. (1996). Geographically weighted regression: a method for exploring spatial non-stationarity. *Geographical Analysis*, 28, 281–298.

[4] Costafreda-Aumedes, S., Vega-Garcia, C., Comas, C. (2018). Improving fire season definition by optimized temporal modelling of daily human-caused ignitions. *Journal of Environmental Management*, 217, 90-99

[5] Schultz, C., Alegría, A.C., Cornelis, J., Sahli, H. (2016). Comparison of spatial and aspatial logistic regression models for landmine risk mapping. *Applied Geography*, 66, 52–63.

[6] Simões, P., Carvalho, M.L., Aleixo, S., Gomes, S. , Natário, I. (2017). A Spatial Econometric Analysis of the Calls to The Portuguese National Health Line, *Econometrics*, 5, 24, MDPI journals.

[7] Simões, P., Carvalho, M.L., Aleixo, S., Gomes, S., Natário, I. (2018). Spatio-Temporal Modelling of the Number of Calls to a National Health Line for Assessing Hospital Savings, *Submitted*.

[8] Sofro, A., Oktaviarina, A. (2018). *Gaussian Process Regression Model in Spatial Logistic Regression* Journal of Physics: Conference Series, 947, 012005

[9] Wu, W., Zhang, L. (2013). Comparison of spatial and non-spatial logistic regression models for modeling the occurrence of cloud cover in north-eastern Puerto Rico. *Applied Geography* 37, 52–62.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

SCREENING PROCEDURES BASED IN MODIFIED CLASSIFICATION TREES APPLIED TO PAEDIATRIC FAMILIAL HYPERCHOLESTEROLEMIA

João Albuquerque[1,2], Ana Catarina Alves[3,4], Mafalda Bourbon[3,4] and Marília Antunes[1]

[1] Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[2] Departamento de Bioquímica da Faculdade de Medicina da Universidade do Porto, 4200-450, Porto, Portugal
[3] Cardiovascular Research Group, Research and Development Unit, Departamento de Promoção da Saúde e Prevenção de Doenças não Transmissíveis, Instituto Nacional de Saúde Doutor Ricardo Jorge, IP, Lisboa, Portugal
[4] Instituto de Biossistemas e Ciências Integrativas– BioISI, Faculdade de Ciências, Universidade de Lisboa, 1749-016 lisboa, Portugal.

ABSTRACT

Familial Hypercholesterolemia (FH) is an autosomal dominant disorder of lipid metabolism, resulting in severe dyslipidemia. Clinical criteria for FH diagnosis are generally based on family history, presence of physical signs, and low density lipoprotein (LDL-c) and total cholesterol (CT) levels, although only the genetic study allows confirming this condition. The high false positive rate presented by clinical diagnosis represents a serious problem, since the correct identification of dyslipidemia etiology is crucial for the assessment of cardiovascular risk and therapeutic approach, particularly in the pediatric age group. Therefore, the main purpose of this work was to develop a classification model for FH based on a modified version of the classic decision tree, using several biochemical markers as predictor variables. The modified decision tree was fitted to sample data, and the correspondent operating characteristics were compared to the ones resulting from Simon Broome (SB) clinical criteria for FH diagnosis. Overall, the modified decision tree model seems to be a good alternative to traditional clinical criteria for FH diagnosis.

Keywords and key sentences: Familial Hypercholesterolemia, Decision trees, Simon Broome criteria, Operating characteristics.

1. INTRODUCTION

Dyslipidemia represents a major risk factor for cardiovascular disease. Familial Hypercholesterolemia (FH) is the most common of all identified monogenic dyslipidemias, and is characterized by elevated plasmatic cholesterol (CT) concentrations, in particular low density

lipoproteins (LDLc)[1]. It is an autosomal dominant pathology, mainly related to mutations in the LDLR gene, and less frequently in APOB and PCSK9 genes [2]. Early diagnosis of FH, especially at pediatric age, has been associated with a significant reduction in cardiovascular disease risk, supporting the introduction of precocious and/or more aggressive therapeutic measures, in a cost-effective process [1,2]. Simon Broome criteria for the diagnostic of FH are among the most frequently used in clinical setting, and are based on CT and LDLc plasmatic concentration above 260 mg/dL and 155 mg/dL respectively, presence of tendinous xanthomas and family history, although only genetic testing can confirm the diagnostic [3].
The Portuguese Study of Familial Hypercholesterolemia (EPHF) has developed the molecular study of this pathology for dyslipidemic patients with clinical criteria for FH, having diagnosed up to 642 FH patients since 1999 [4]. However, when considering the conservative prevalence estimate of 1:500 individuals [5], we realize this pathology is severely under diagnosed in our country, having identified only around 3,2% of FH carriers. Moreover, among the patients that have been referred for molecular study, only around 42% have revealed a positive diagnostic for FH [4]. This high false positive rate represents a heavy burden in terms of healthcare costs, limiting the access to the genetic study of a larger universe of potential FH cases, at an earlier stage.
The present work is part of an ongoing current project focusing in the improvement of the ability to screen potential FH cases based on different biochemical indicators, providing an alternative to the currently used Simon Broome criteria. For this purpose, we will use decision tree-based models, namely, a modified decision tree implementation method, which will sequentially eliminate predictor variables, as they are introduced in the model. The biochemical panel will include CT, LDLc, HDL cholesterol (HDLc) and triglycerides (TG) concentrations. The resulting decision tree will be compared with Simon Broome clinical criteria in terms of accuracy and efficiency. These are the very first results in this project. In a subsequent stage, a larger number of easy to obtain variables will be considered as candidates, as well as the physical signs and family history (considered in Simon Broome criteria). Also, bootstrap resampling techniques, inspired in the random forests approach will be used in an atemptive to improve the discriminant power of the classifier.

2. METHODS

2.1. The Sample

The sample used in this study is constituted by participants in the EPHF, previously presented. A total of 349 cases of both sexes, at pediatric age (2 to 17 years), meeting the clinical criteria for dyslipidemia [6], have been included.

2.2. Biochemical and molecular analysis

Blood samples for DNA extraction and biochemical panel determination were collected. Genomic DNA was extracted using a Wizard® Genomic DNA Purification Kit (Promega) according to manufacturer's instructions. Lipidic parameters, specifically CT, LDLc, HDLc and TG were assessed using a Cobas Integra 400 Plus (roche) analyzer, and measured in mg/dL [7]. For the molecular study, mutations were searched at specific sites in the LDLR, APOB ans PCSK9 genes, through fragment amplification by polymerase chain reaction (PCR), followed by direct sequencing using Sanger's method. Gene rearrangements in LDLR gene were also searched through Multiplex Ligation-dependent Probe Amplification (MLPA) technique [7]. Cases of polygenic mutation, unknown alteration in one of the referred genes or homozygous FH were excluded from the study. In case no molecular alteration was observed, the participant was classified as negative for FH.

2.3. Statistical procedures

Entropy and information gain measures were calculated to quantify the ability of each of the biomarkers to discriminate between presence and absence of FH, and biomarkers were

ranked accordingly [8]. A modified version of the decision tree method was posteriorly implemented. The main difference in comparison to the traditional approach consists in the sequential exclusion of predictor variables as they are used in each tree node. In the following nodes, entropy measures are calculated for the remaining variables, and the one with highest information gain is selected, repeating this procedure throughout the tree. The final tree was pruned using C5 algorithm to avoid overfitting [9]. The motivation behind this approach was to produce a classification rule that would resamble typical medical criteia, which usually consider single cutpoints for biomarkers. The modified decision tree model was then compared with the biochemical markers used in Simon Broome criteria for FH diagnosis. A confusion matrix was built for both models, by comparison with molecular study results, and different operating characteristics were used for performance comparison. These included accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) [10]. Statistical analysis was performed using R and R Studio.

3. RESULTS

The biomarkers used were TC, LDLc, HDLc, and TG. The resulting decision tree can be observed in Fig. 1.

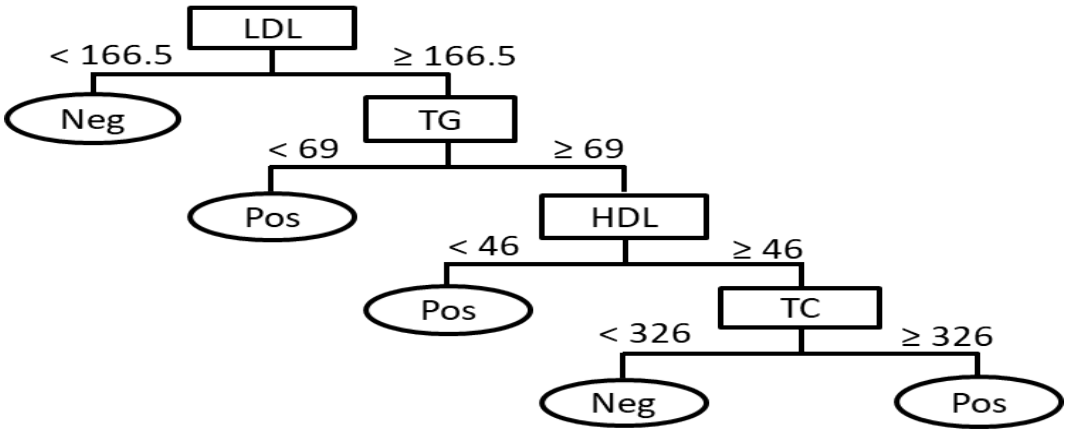


Figure 1: Estimated decision tree.

Unlike the Simon Broome criteria, which use a defined cut-off value to classify the patient as FH positive or negative, the decision tree uses a sequential approach for this effect. The most informative variable was still LDLc, which makes sense since this disease primary affects LDLc metabolism. While patients with LDLc$<$ 166.5 are immediately classified as negative in the pruned tree, for patients with LDLc$\geq$ 166.5 the tree suggests TG as the following reference variable. At this node, patients with TG$<$ 69.0 are classified as positive. This can be understood from the stand point that high TG levels may be more related to an environmental dyslipidemia than FH. TC comes at the end of the tree, which can be explained by the fact that this marker is in fact a linear combination of LDLc and HDLc fractions, both appearing in nodes above. The modified version of the decision tree allows simplifying its interpretation, since each predictor variable is only used once. The confusion matrix for decision tree and SB models, and respective operating characteristics are presented in Table 1. Decision tree-based model shows increased accuracy, specificity and PPV compared to Simon Broome criteria. This means that the decision tree model correctly classifies the patient more often than Simon Broome criteria on one hand, and also that is has better ability to exclude negative cases. On the other side, Simon Broome criteria present better sensitivity and NPV, suggesting higher ability to retain positive cases. Given the conservative cut-off values of 155 mg/dL for LDL and 260 mg/dL for TC, this may be accomplished at the expense of retaining a high number

					Operating Characteristics				
method	TP	FP	TN	FN	Accuracy	Sensitivity	Specificity	PPV	NPV
DT	75	9	163	19	0.895	0.798	0.948	0.893	0.896
SB	91	68	104	3	0.733	0.968	0.605	0.572	0.972

Table 1: Number of TP, FP, TN and FN, and operating characteristics (OC) for the decision tree (DT) and Simon Broome (SB) criteria.

of false positive cases (68 false positives, against 9 for the tree-based model), which can prove to be costly and inefficient in clinical practice.

4. CONCLUSIONS

Overall, the modified decision tree shows increased accuracy and specificity compared to SB criteria. Increased sensitivity in SB model is accomplished at the expense of high false positive retention, which is likely to have an adverse effect in the cost-effectiveness relation. By avoiding the repetition of predictor variables in this modified version we expect to simplify the model interpretation for physicians and other health professionals. Although validation of these results through bootstrap resampling techniques is still in process, decision tree classification methods seem to be a viable alternative for FH diagnosis.

ACKNOWLEDGMENTS

The research by João Albuquerque was supported by the programmme Norte2020 (operação NORTE-08-5369-FSE-000018); The research by Marília Antunes was supported by national funds through FCT under the project UID/MAT/00006/2013.

References

[1] Goldberg, A.C., Hopkins, P.N., Toth, P.P., Ballantyne, C.M., Rader, D.J., Robinson, J.G., et al (2011). Familial hypercholesterolemia: Screening, diagnosis and management of pediatric and adult patients: Clinical guidance from the National Lipid Association Expert Panel on Familial Hypercholesterolemia. *J Clin Lipidol*; 5(3 Suppl):S1-8;

[2] Watts,G.F., Gidding,S., Wierzbicki,A.S., et al (2014). Integrated guidance on the care of familial hypercholesterolaemia from the International FH Foundation. *Int J Cardiol*; 171:309–325;

[3] Scientific Steering Committee on behalf of the Simon Broome Register (1991). Risk of fatal coronary heart disease in familial hypercholesterolaemia. *BMJ*; 303:893-896;

[4] Medeiros,A.M., Alves, A.C., Francisco,V., Bourbon, M. (2010). Update of the Portuguese Familial Hypercholesterolaemia Study. *Atherosclerosis*; 212:553–558;

[5] Nordestgaard, B.G., Chapman, M.J., Humphries, S.E., Ginsberg, H.N., Masana, L., Descamps, O.S., et al (2013). Familiar hypercholesterolemia is underdiagnosed and untreated in general population; guidance for clinicians to prevent coronary heart disease. *Eur Heart J*;34:3478–90;

[6] Jolliffe, C.J., Janssen, I. (2006). Distribution of lipoproteins by age and gender in adolescents. *Circulation*; 114:1056–62

[7] Benito-Vicente, A., Alves, A.C., Etxebarria, A., Medeiros, A.M., Martin, C., Bourbon, M. (2015). The importance of an integrated analysis of clinical, molecular, and functional data for the genetic diagnosis of familial hypercholesterolemia. *Genetics in Medicine*; 17(12), 980;

[8] Kingsford, C., Salzberg, S.L. (2008). What are decision trees? *Nat Biotechnol*, 26(9): 1011–1013

[9] Berry, M; Linoff, G (2011). *Data mining techniques for marketing, sales and customer support*, 2nd Ed. Wiley

[10] Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

MEDIDAS INFORMATIVAS EM ESTUDOS LONGITUDINAIS: UM ESTUDO DE SIMULAÇÃO

Adriana Vieira[1] e Inês Sousa[1]

[1]Departamento de Matemática e Aplicações da Universidade do Minho

RESUMO

Um dos grandes problemas em estudos cujos dados são de natureza longitudinal é a existência de tempos de observação informativos. Estes tempos de observação são caracterizados pela sua relação com a variável resposta, contrariamente ao que ocorre noutros estudos longitudinais, onde variável resposta e tempo são independentes (ou independentes dado um conjunto de covariáveis). Na presença de tais medidas informativas, o uso das metodologias utilizadas no caso da independência ser válida conduz a estimadores enviesados e, consequentemente, conclusões incertas.
Neste trabalho pretendemos apresentar alguns modelos alternativos, que se enquadrem na problemática, demonstrando as suas diferenças através de um estudo de simulação.

Palavras-chave: dados longitudinais, medidas informativas, simulação, tempos de observação

1. INTRODUÇÃO

A análise de dados longitudinais representa um papel fundamental numa multiplicidade de áreas distintas, nomeadamente na medicina. Uma das grandes dificuldades neste tipo de estudo prende-se com diferentes tempos de observação para diferentes indivíduos, tempos estes que são tratados como independentes da variável resposta.
Uma dificuldade ainda maior ocorre quando os diferentes tempos de observação estão relacionados com a variável resposta. Por exemplo, o médico decide marcar mais, ou menos consultas de acordo com o estado de saúde do paciente.
Exemplos de dados longitudinais com medidas informativas podem ser encontrados em diversos estudos. Em [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**] são usados dados relativos à incidência de cancro da bexiga - alguns pacientes foram observados com muito mais regularidade que outros o que parece relacionado com a ocorrência dos tumores. Noutros estudos encontra-se um conjunto de dados relativos aos efeitos cardiotóxicos da quimioterapia com doxorrubicina em leucemia linfoblástica aguda infantil - pacientes com medidas anormais da massa ventricular esquerda são passíveis de maior número de consultas médicas [**?**, **?**]. Outra base de dados usual prende-se com dados relativos à falha cardíaca crónica - pacientes cujas visitas são mais frequentes tendem a pagar mais pelos cuidados médicos, além de terem uma taxa de mortalidade maior [**?**, **?**].

Nos casos em que tempos de observação e variável resposta se encontram relacionados, como nos casos acima, uma simples análise longitudinal produzirá estimadores enviesados [**?**] e, consequentemente, conclusões incertas. Assim sendo, passa a ser necessário o desenvolvimento de novas metodologias, que permitam a inclusão desta característica.
Pretendemos então apresentar aqui alguns modelos alternativos, que se enquadrem na problemática, demonstrando as suas diferenças através de um estudo de simulação

2. MODELO LONGITUDINAL COM MEDIDAS INFORMATIVAS

Consideremos um estudo longitudinal com uma amostra aleatória de n indivíduos. Seja Y_{ij} a variável resposta de interesse medida no indivíduo i e tempo j e T_{ij} o tempo da medida j para o indivíduo i, $i = 1, ..., n$, $j = 1, .., m_i$. Seja $W(s)$ um processo Gaussiano estacionário, não observável e contínuo no tempo, medido em $s = (s_1, ..., s_M)$ com $M \in \mathbb{R}_0^+$. Vamos assumir então dados gerados de acordo com

$$[Y_{ij}|W(T_{ij}), T_{ij}] \sim W(T_{ij}) + Z_{ij}$$
$$[\lambda(T_{ij})|W_{historico}(s)] \sim exp\{\mathcal{F}(W_{historico}(s))\}$$

onde $Z_{ij} \sim N(0, \tau^2)$ e $\lambda(s)$ representa a intensidade das visitas em s. A função de intensidade pode depender do histórico da variável resposta.
Sejam $(t_{i1}, ..., t_{im_i})$ um subconjunto de $(s_1, ..., s_M)$. Neste estudo vamos considerar os seguintes modelos para a intensidade:

1. $\lambda(t_{ij}) = exp\{\alpha + \beta W(t_{ij})\}$
2. $\lambda(t_{ij}) = exp\{\alpha + \beta[W(t_{ij}) - W(t_{ij-1})]\}$
3. $\lambda(t_{ij}) = exp\left\{\alpha + \beta \sum_{s=0}^{t_{ij}} W(t_{ij}) w(t_{ij} - s)\right\}$, onde $\sum_{s=0}^{t_{ij}} w(t_{ij} - s) = 1$.

3. ESTUDO DE SIMULAÇÃO

Recorrendo a um processo de simulação é possível gerar $Y_{ij} = Y(T_{ij})$. Tal processo implica os seguintes passos:

1. Gerar o processo não observado $W(s)$.

   Note-se que apesar de contínuo no tempo, para a simulação propriamente dita, o processo $W(s)$ será avaliado tomando o tempo como discreto, ou seja, $(s_1, ..., s_M)$ será um conjunto discreto de pontos equidistantes, também estes simulados.

   Assumindo um caso particular, o processo $W(s)$ é gerado como $E[W(s)] = 0$, $Var(W(s)) = \sigma^2$ e $Corr(W(s_k), W(s_l)) = exp\left\{-\frac{1}{\phi}|s_l - s_k|\right\}$, sendo ϕ o parâmetro chamado *range*.

2. Simular o vetor dos tempos t_{ij} tendo em conta cada uma das intensidades apresentadas na secção 2.

   Em cada ponto s_q, $q = 1, ..., M$ é calculada a probabilidade de se ter uma observação, isto é, a probabilidade de s_q ser um tempo de observação t_{ij}.

   Note-se que a probabilidade de sucesso p, isto é, a probabilidade de s_q ser um tempo de observação t_{ij} é o parâmetro de uma distribuição Geométrica, cuja variável aleatória seria "número de falhas até s_q ser t_{ij}". Assim, assumindo uma distribuição Exponencial usual de parâmetro $\lambda > 0$ e conhecida a relação entre as distribuições Geométrica e Exponencial tem-se $p = 1 - exp\{-\lambda\}$[**?**]. Neste caso λ é a nossa intensidade.

3. O processo Y_{ij} é gerado diretamente como $Y_{ij} = Y(T_{ij}) = W(T_{ij}) + Z_{ij}$.

4. CONCLUSÕES

Vejamos alguns resultados tomando como intensidade o modelo 1, isto é,

$$\lambda(T_{ij})|W_{historico}(s) = exp\{\alpha + \beta W(t_{ij})\}.$$

Consideramos $\phi = 0.25$, $\sigma^2 = 1$ e $\tau^2 = 0$. Definimos ainda um total de 201 pontos temporais equidistantes, a ser gerados, com $M = 20$ meses.
Note-se que assumindo $\tau^2 = 0$, $Y_{ij} = W(T_{ij})$, sendo portanto natural o que acontece na Figura 1 - o processo Y_{ij} coincide com o processo $W(T_{ij})$.
A Figura 1 permite perceber que diferenças nos parâmetros da intensidade resultam em diferenças significativas no processo longitudinal. Sabendo que, se $\beta = 0$ o processo não é mais que um caso longitudinal usua, é perceptível pelas diferenças nas Figuras 1(d) e 1(e), que ignorar o histórico do processo pode levar a erros.
Uma vez que considerar o modelo 1 de intensidade pode ser visto como um dos casos mais simples, e, atendendo aos resultados obtidos, podem também esperar-se disparidade usando modelos mais complexos (2 a 5). Assim, pode concluir-se a importância de estudar esta característica - medidas temporais informativas.

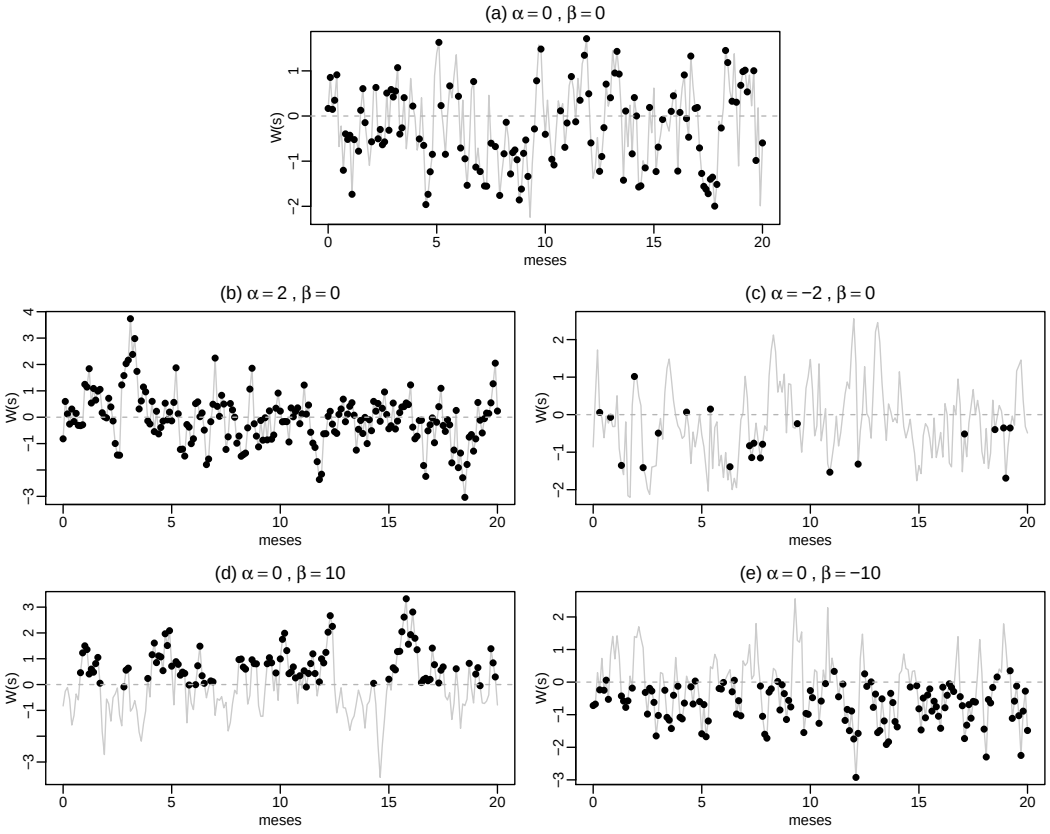


Figura 1: Processo observado ao longo do tempo, com intensidade $exp\{\alpha + \beta W(t_{ij})\}$ onde: $(\alpha, \beta) = (0, 0)$ em (a); $(\alpha, \beta) = (2, 0)$ em (b); $(\alpha, \beta) = (-2, 0)$ em (c); $(\alpha, \beta) = (0, 10)$ em (d); $(\alpha, \beta) = (-10, 0)$ em (e). Pode ver-se que aumentando (respetivamente diminuindo) o valor de α a intensidade aumenta (respetivamente diminui), relativamente a um valor de α nulo. Aumentando (respetivamente diminuindo) o valor de β o processo toma valores maioritariamente positivos (respetivamente negativos), relativamente a um valor de β nulo.

AGRADECIMENTOS

Este trabalho teve o apoio financeiro através da bolsa de doutoramento 128191/2016 pela FCT I.P., do Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) ao primeiro autor.

Referências

[1] Zhang Y. (2002). A semiparametric Pseudolikelihood Estimation Method for Panel Count Data. *Biometrika* 89, 39–48.

[2] Sun J., Park D., Sun L., Zhao X. (2005). Semiparametric Regression Analysis of longitudinal Data With Informative Observation Times. *Journal of the American Statistical Association* 100, 882–889.

[3] Sun J., Sun L., Liu D. (2007). Regression Analysis os Longitudinal Data in the Presence of Informative Observation and Censoring Times. *Journal of the American Statistical Association* 102, 1397–1406.

[4] Song X., Mu X., Sun L. (2012). Regression Analysis of Longitudinal Data with Time-dependent Covariates and Informative Observation Times. *Scandinavian Journal of Statistics* 39, 248–258.

[5] Zhao X., Tong X., Sun L. (2012). Joint Analysis of Longitudinal Data with Dependent Observation Times. *Statistica Sinica* 22, 317–336.

[6] Fang S., Zhang H., Sun L. (2016). Joint analysis of longitudinal data with additive mixed effect model for informative observation times. *Journal of Statistical Planning And Inference* 169, 43–55.

[7] Du T., Ding J., Sun L. (2016). Joint modeling and estimation for longitudinal data with informative observation and terminal event times. *Communications in Statistics - Theory and Methods* 45, 6521–6539.

[8] Miao R., Chen X., Sun L. (2016). Analyzing Longitudinal Data with Informative Observation and Terminal Event Times. *Acta Mathematicae Applicatae Sinica, English Edition* 32, 1035–1052.

[9] Pei Y., Du T., Sun L. (2016). Time-varying latent model for longitudinal data with informative observation and terminal event times. *Sci China Math* 59, 1–18.

[10] Lipsitz S., Fitzmaurice G., Ibrahim J., Gelber R., Lipshultz S. (2002). Parameter Estimation in Longitudinal Studies with Outcome-Dependent Follow-Up. *Biometrics* 58, 621–630.

[11] Fitzmaurice G., Lipsitz S., Ibrahim J., Gelber R., Lipshultz S. (2006). Estimation in regression models for longitudinal binary data with outcome-dependent follow-up. *Biostatistics* 7, 469–485.

[12] Fitzmaurice G., Lipsitz S., Ibrahim J., Gelber R., Lipshultz S. (2006). Estimation in regression models for longitudinal binary data with outcome-dependent follow-up. *Biostatistics* 7, 469–485.

[13] Han M., Song X., Sun L., Liu L. (2014). Joint modeling of longitudinal data with informative observation times and dropouts. *Statistica Sinica* 24, 1487–1504.

[14] Lin H., Scharfstein D., Rosenheck R. (2004). Analysis of longitudinal data with irregular outcome-dependent follow-up. *J. R. Statist. Soc. B* 66, 791–813.

[15] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications, Volume II*. John Wiley & Sons, New York.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

ENSAIOS CLÍNICOS: HISTÓRIA E EVOLUÇÃO

Raquel Correia[1], Fernanda Diamantino[2]

[1]Faculdade de Ciências da Universidade de Lisboa
[2]Faculdade de Ciências da Universidade de Lisboa e CEAUL

RESUMO

A história e a evolução dos ensaios clínicos é extensa e fascinante. O primeiro ensaio registado (562 a.C.) lida com alimentação e é descrito num dos livros da Bíblia. Até aos RCT (*Random Clinical Trials*) atuais, a história é longa passando por três desafios principais: desenvolvimento científico, procedimentos éticos e regulamentação.

Palavras-chave: ensaios clínicos, história, ética.

1. PRIMEIRO ENSAIO CLÍNICO (562 a.C.)

Nabuconodosor, rei da Babilónia, pediu que escolhessem jovens israelitas exilados, da realeza e da nobreza, para servir na corte [1]. O rei deu ordens para que a comida e o vinho que lhes eram servidos fossem os mesmos da casa real.
Durante três anos os jovens iam ser preparados para entrar ao serviço do rei. Alguns desses jovens fizeram questão em se manter fiéis às regras de alimentação do seu povo. Um deles pediu ao encarregado que durante dez dias só lhes dessem legumes e água.
Após esses dias verificou-se que os jovens tinham um aspeto mais saudável e robusto em comparação com os que comiam segundo a ementa real. Após esta experiência foi-lhes permitido continuar a dieta. Deste modo, esta experiência tornou-se o primeiro ensaio clínico.

2. AMBROISE PARÉ E O TRATAMENTO DE FERIMENTOS (1537)

Em 1537, França estava em guerra. O cirurgião Ambroise Paré fez a sua primeira descoberta médica no tratamento de ferimentos de balas quando se encontrava num cerco.
Este tipo de ferimento era tratado normalmente por cauterização com óleo a ferver. Devido ao cerco, Paré ficou sem óleo e teve que optar por fazer o tratamento com gema de ovo, óleo de rosas e terebentina e notou melhoria nos ferimentos.
Após esta descoberta, Paré defendia que se devia evitar a cauterização, visto que, para além de não causar sofrimento ao paciente, no local do ferimento a pele não ficaria com marcas de queimaduras.
Depois de Paré, passar-se-iam quase 200 anos até surgir outro ensaio clínico controlado.

3. JAMES LIND E O ESCORBUTO (1747)

A 20 de maio de 1747 no navio real *Salisbury*, que tinha o propósito de controlar o tráfico marítimo, um décimo da tripulação tinha escorbuto [3]. Desses indivíduos, o cirurgião James Lind escolheu 12 e aplicou 6 tipos de terapia: sidra, "elixir de vitríolo" feito a partir de ácido sulfúrico, vinagre, água do mar, laranjas e limões e uma mistura picante à base de mostarda. No fim de maio, os dois homens a quem foram dados os citrinos, já estavam quase curados. Como os citrinos na altura eram muito caros, James Lind hesitou em recomendar os citrinos para o tratamento do escorbuto.
Lind publicou os seus resultados em 1753, mas a introdução de sumo de limão ou laranja na dieta dos marinheiros britânicos só surgiu cerca de quatro décadas mais tarde.

4. APARECIMENTO DO PLACEBO (1772)

Em 1772, um médico e farmacologista escocês, William Cullen acrescentou a palavra placebo ao vocabulário médico, mas o crédito foi dado a um médico inglês, Alexander Sutherland. Em notas manuscritas William Cullen decreve que administrou um placebo "puro" para confortar um paciente que estava prestes a morrer.
Uma das principais razões da integração do placebo foi a necessidade de satisfazer as exigências e expetativas dos pacientes. Outro motivo foi o facto de os pacientes insistirem em ter um medicamento, o que levava a ser-lhes dado um medicamento inerte: os pacientes ficavam satisfeitos e o tratamento não produzia qualquer efeito.
No século XVIII os médicos não proviam os pacientes com placebos "puros", mas forneciam medicamentos com pouco efeito. Robert Hooper, em 1811, cita o placebo como "o nome dado a qualquer medicamento administrado mais para agradar do que beneficiar o paciente" [1].

5. CONTROLO COM DUPLA OCULTAÇÃO (1943)

Durante a 2.ª Guerra Mundial, o *Medical Research Council* começou a investigar o uso de patulina [1] para o tratamento de constipações. Neste comité encontravam-se Sir Harold Himsworth e os estatísticos M. Greenwood e W. J. Martin.
Este estudo envolveu cerca de 100 indivíduos com constipação que trabalhavam em escritórios e fábricas. A experiência era controlada rigorosamente com o propósito de manter tanto o médico como o paciente "cegos" em relação ao tratamento. A enfermeira assistente preencheu os formulários separadamente e descolou o código do rótulo do frasco que continha ou não o medicamento antes de convocar o paciente para ser visto pelo médico.
Os estatísticos consideraram a organização da experiência eficiente, mas o ensaio não mostrou que a patulina tivesse um efeito protetor da doença.

6. INTRODUÇÃO DA ALEATORIDADE (1948)

A ideia de aleatoriedade nos ensaios clínicos foi introduzida em 1923. O *Medical Research Council* para investigar o uso da estreptomicina no tratamento da tuberculose integrava Sir Geoffrey Marshall e os estatísticos Sir Austin Bradford Hill e Philip Hart [2].
A investigação iniciou-se em 1947. Devido ao facto da quantidade do antibiótico nos Estados Unidos estar, na altura, limitada, não haveria problemas éticos se os pacientes, que sofriam de tuberculose, não fossem tratados com estreptomicina.
Esta experiência foi um bom exemplo de um tratamento rigoroso e bem implementado. Durante o procedimento foram tomadas medidas para que exames, tais como os raios-x e a sua interpretação, fossem realizados por médicos que não sabiam que tipo de tratamento o paciente estava a receber.

[1] A patulina surgiu em 1940 após a descoberta da penicilina, tendo sido utilizada como spray nasal para tratamento de constipações e como pomada para tratamento antifúngico da pele.

Bradford Hill, durante vários anos, criou um método para este tipo de experiências, mas só o tinha aplicado na prevenção de doenças. Com esta experiência conseguiu finalmente usar a sua metodologia num ensaio onde o objetivo era curar uma doença. Deste modo, conseguiu instituir a aleatorização nos ensaios clínicos.

7. A EVOLUÇÃO ÉTICA E REGULAMENTAR

A origem da ética em experiências vem desde a criação do juramento de Hipócrates onde os médicos juravam sempre evitar o sofrimento do paciente. Mas, como o passado comprova, nem todos os médicos cumpriam o juramento como, por exemplo, nas experiências realizadas em seres humanos pelos nazis.
A FDA (*Food and Drug Administration*) foi fundada em 1862 como uma instituição científica. Após o Congresso dos Estados Unidos ter aprovado o *Food and Drugs Act*, em 1906, tornou-se numa organização que procurou aplicar o legislado.
A legislação a partir do *Food and Drugs Act*, progressivamente, exigiu maior responsabilidade ao comércio, tanto de alimentação como de medicamentos, o que teve como consequência o aumento de ensaios clínicos para testar a segurança e a eficácia de medicamentos.
Em 1947 foi criado o Código de Nuremberga, o primeiro guia internacional de ética em pesquisa médica que envolvia seres humanos. Este código deu ênfase à necessidade de os pacientes serem voluntários nas experiências.
Em 1948, a Declaração Universal dos Direitos do Homem, expressou preocupação em relação aos seres humanos que estariam a ser maltratados involuntariamente. Assim, esta declaração tornou-se um documento fundador da garantia do respeito pela pessoa humana.
Em 1964, em Helsínquia, a Associação Médica Mundial, concebeu a Declaração de Helsínquia que foi o primeiro código de conduta para a investigação médica no homem.
No artigo 7.º do Pacto Internacional sobre os Direitos Civis e Políticos, escrito em 1966, está referido que "Ninguém poderá ser submetido a torturas, penas ou tratamentos cruéis, desumanos ou degradantes. Em particular, ninguém será submetido sem o seu livre consentimento a experiências médicas ou científicas". Este pacto foi assinado por Portugal a 7 de outubro de 1976.
Em 1996, a Conferência Internacional de Harmonização publicou as Boas Práticas Clínicas, que se tornaram o padrão universal para a conduta ética nos ensaios clínicos.

8. ENSAIOS CLÍNICOS EM PORTUGAL

Em dezembro de 1988 foi criado em Coimbra o Centro de Estudos de Bioética. Mais tarde, em junho de 1990, tornar-se-ia no Conselho Nacional de Ética para as Ciências da Vida (CNECV) com o intuito de definir as grandes directrizes no campo da bioética [4].
O Decreto-Lei n.º 97/94, de 9 de abril, estabelece, em 1994, "as normas a que devem obedecer os ensaios clínicos a realizar em seres humanos", que só se tornou exequível com a publicação do Decreto-Lei n.º 97/95, de 10 de maio, que definiu a composição e funcionamento das Comissões de Ética para a Saúde (CES), uma por cada instituição.
A Lei n.º 46/2004, de 19 de agosto, criou a Comissão de Ética para a Investigação Clínica (CEIC) a qual, em 2005, se tornou a autoridade competente que emite parecer sobre a realização de ensaios clínicos com medicamentos de uso humano.
A CEIC, segundo a Lei n.º 21/2014, de 16 de abril, é "um organismo independente constituído por individualidades ligadas à saúde e a outras áreas de atividade, cuja missão principal é garantir a proteção dos direitos, da segurança e do bem-estar dos participantes nos estudos clínicos, através da emissão de um parecer ético sobre os protocolos de investigação que lhe são submetidos".
O Infarmed (Autoridade Nacional do Medicamento e Produtos de Saúde I.P.) foi criado em

1993, com o intuito de monitorizar as atividades relacionadas com os medicamentos. Atualmente é quem autoriza os estudos e monitoriza a segurança da utilização de medicamentos experimentais nos ensaios clínicos, garantindo que os mesmos ocorrem de acordo com a legislação aplicável.

Referências

[1] Bhatt, A. (2010). Evolution of Clinical Research: A History Before and Beyond James Lind. *Perspect Clin Res.* Jan-Mar, 1(1), 6–10.

[2] Collier, R. (2009). Legumes, lemons and streptomycin: A short history of the clinical trial. *CMAJ* 180(1)), 23–24.

[3] Suttom, G. (2003). Putrid gums and "Dead Men's Cloaths": James Lind aboard the *Salisbury*. *J R Soc Med* 96, 605–608.

[4] Neves, M.C.P. (2016). *A Origem da Bioética em Portugal Através dos Seus Pioneiros*. Fronteira do Caos, Porto.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

TEMPERATURA À SUPERFÍCIE DO MAR E ÍNDICE DE AFLORAMENTO COSTEIRO: MODELAÇÃO E COMPARAÇÃO AO LONGO DA COSTA DE PORTUGAL CONTINENTAL

Bruno Monteiro[1], M. Rosário Ramos[1] e Clara Cordeiro[2]

[1]Universidade Aberta, Portugal
[2] Faculdade de Ciências e Tecnologia da Universidade do Algarve, e Centro de Estatística e Aplicações, Universidade de Lisboa, Portugal

RESUMO

O aquecimento global, e em particular o dos oceanos, é um assunto de extrema importância devido aos seus efeitos e à necessidade de adaptação das populações face a este fenómeno. As camadas superficiais dos oceanos têm um papel importante na manutenção da temperatura global, contudo, em várias regiões planetárias, a temperatura à superfície do mar (SST- do inglês *Sea Surface Temperature*) tem sofrido um aumento significativo nas últimas décadas (Castro, 2003). A análise da SST permite estudar a evolução deste aumento na medida em que se explora modelos para o descrever. Neste trabalho é estudado o comportamento da SST no periodo entre janeiro de 1982 a dezembro de 2013, nas regiões Norte, Centro e Sul de Portugal continental. Em cada uma das regiões foram considerados os dados de satélite obtidos pelo *E.U. Copernicus Marine Service Information* em duas localizações: próxima da costa (distância inferior a 20 km da costa), e afastada da costa (entre 135 e 150 km). Foi igualmente objeto de interesse o fenómeno denominado por afloramento costeiro (*Upwelling*), que é um fenómeno recorrente na costa portuguesa e responsável pelas temperaturas superficiais do mar mais baixas nos meses mais quentes do ano (em pleno verão). Neste contexto é proposto o Índice de Afloramento Relativo (IAR) que possibilita uma interpretação, em termos relativos, da variação da temperatura devida a este fenómeno e ainda, estabelecer uma comparação entre as regiões em estudo. Para cada região, foram utilizados modelos de séries temporais para a série IAR, nos quais se incluem os modelos de Decomposição Clássica e de Holt-Winters, e ainda os modelos ARIMA sobre a série residual, removida uma tendência linear e a componente sazonal. A seleção do melhor modelo foi auxiliada por medidas de ajustamento. Foi possível, neste estudo, identificar um padrão sazonal anual; foi ainda possível observar que o efeito do afloramento costeiro é mais evidente nos meses mais quentes do ano, em que, devido a este fenómeno, o aumento da SST é mais evidente nas localizações distantes da costa relativamente às localizações costeiras com a mesma latitude.

Palavras-chave: Afloramento costeiro, Temperatura à superfície do mar, Índice de Afloramento Relativo, modelos ARIMA, Decomposição Clássica, Holt-Winters.

AGRADECIMENTOS

Clara Cordeiro research has been supported by FCT - Fundação para a Ciência e a Tecnologia, through the project UID/MAT/00006/2013 (CEA/UL)

Referências

[1] Castro, Peter (2003). *Marine Biology*, 4th edition. U.S.A.: McGraw-Hill Companies, ISBN: 978-0-072-85290-5.

[2] Goela, Priscila Costa; Cordeiro, Clara; Danchenko, Sergei; Icely, John; Cristina, Sónia; Newton, Alice (2016). Time series analysis of data for sea surface temperature and upwelling components from the southwest coast of Portugal.*Journal of Marine Systems* 163, 12–22.

[3] Gonzalez-Nuevo, Gonzalo; Gago, Jesus; Cabanas, Jose M. (2014). Upwelling index: a powerful tool for marine research in the NW Iberian upwelling system. *Journal of Operational Oceanography*, 7, Issue 1, 47–57.

[4] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[5] Shumway, Robert H.; Stoffer, David S. (2014). *Time Series Analysis and its Applications, with R Examples* EZ, third edition, Free texts in Statistics ISBN: 978-1-4419-7865-3.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

MODELAÇÃO DE VALORES EXTREMOS DE TENSÃO ARTERIAL

Constantino Pereira Caetano[1] e Patricia de Zea Bermudez[2]

[1] Faculdade de Ciências da Universidade de Lisboa

[2] CEAUL e Faculdade de Ciências da Universidade de Lisboa

RESUMO

A hipertensão é considerada um fator de risco das doenças cardiovasculares [3]. Em 2005, a Associação Nacional das Farmácias, através do seu Departamento de Cuidados Farmacêuticos, promoveu uma Campanha para identificação de indivíduos suspeitos de risco cardiovascular.

O presente estudo centra-se na análise da hipertensão sistólica isolada, a qual é caracterizada por níveis de pressão arterial sistólica superiores ou iguais a 140 mmHg e de pressão arterial diastólica inferiores a 90 mmHg (actuais *guidelines* da Sociedade Portuguesa de Hipertensão). Este tipo de doença é especialmente frequente em indivíduos idosos [2]. Dado o interesse do estudo residir na análise dos valores elevados de tensão arterial sistólica de indivíduos que apresentam valores normais de tensão arterial diastólica, recorreu-se a teoria de valores extremos. Os dados de tensão arterial sistólica foram analisados por meio do método *Peaks over Threshold* (POT) que consiste em ajustar uma distribuição de Pareto generalizada (GPD) aos excessos ou excedências acima de um limiar suficientemente elevado [5]. Na escolha do threshold foram utilizados métodos clássicos, tais como a representação gráfica da função de excesso médio, e também um método bayesiano [4]. Os modelos ajustados permitem estimar quantis extremos de tensão arterial sistólica e probabilidades de cauda. O estudo foi realizado a nível global e também a nível de distrito. Estudos anteriores destes dados trataram o fator de risco colesterol total [1].

Palavras e frases chave: Teoria de valores extremos, distribuição de Pareto generalizada, hipertensão sistólica isolada, estatística bayesiana.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pela FCT - Fundação para a Ciência e a Tecnologia, Portugal, através do projecto UID/MAT/00006/2013.

Referências

[1] de Zea Bermudez, P. & Mendes, Z. (2012): Extreme Value Theory in Medical Sciences: Modeling Total High Cholesterol Levels, *Journal of Statistical Theory and Practice* 6, 468-491.

[2] Gonzaga, C.C., Sousa M.G. & Amodeo C. (2009). Fisiopatologia da hipertensão sistólica isolada. *Revista Brasileira de Hipertensão* 16, 10-14.

[3] Hajar, R. (2016). Framingham Contribution to Cardiovascular Disease. *Heart Views: The official Journal of the Gulf Heart Association* 17, 78-81.

[4] Lee, J., Fan, Y. & Sisson, S. A. (2015). Bayesian threshold selection for extremal models using measures of surprise. *Computational Statistics & Data Analysis* 85, 84-99.

[5] Pickands III, J. (1975). Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics* 3, 119–131.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

UM CONTRIBUTO DA ANÁLISE ESTATÍSTICA NA GESTÃO DE UMA ESTAÇÃO DE TRATAMENTO DE ÁGUAS RESIDUAIS (ETAR)

A. Manuela Gonçalves[1], M. Teresa Amorim[2] e Marco Costa[3]

[1] CMAT – Centro de Matemática, DMA - Departamento de Matemática e Aplicações, Universidade do Minho, mneves@math.uminho.pt

[2] 2C2T – Centro de Ciência e Tecnologia Têxtil, DET – Departamento de Engenharia Têxtil, Universidade do Minho, mtamorim@det.uminho.pt

[3] CIDMA - Centro de Investigação e Desenvolvimento em Matemática e Aplicações, Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro, marco@ua.pt

RESUMO

A deterioração progressiva dos recursos hídricos e a grande quantidade de água poluída gerada pelas sociedades modernas faz com que as estações de tratamento de águas residuais (ETAR) tenham uma extrema importância na prevenção e controlo da qualidade da água. Numa ETAR, o processo de lamas activadas é a tecnologia mais comumente usada para remover poluentes orgânicos das águas residuais (por meio de suspensão de biomassa bacteriana). Esta é a tecnologia com melhor custo-benefício, é muito flexível e pode ser adaptada a diferentes tipos de águas residuais. Por conseguinte, é muito importante compreender e modelar os processos biológicos utilizados numa ETAR, por forma a estimar os custos de tratamento com maior precisão e estabelecer uma melhor relação custo-eficácia. Neste trabalho a discussão centra-se no estabelecimento de análises e modelos estatísticos a fim de quantificar e caracterizar padrões de interacção entre as águas residuais dos afluentes que são tributários para as ETAR, as variáveis hidro-meteorológicas, as variáveis físico-químicas e as variáveis de custos associados aos tratamentos. O procedimento de modelação estatística foi aplicado a um conjunto de ETAR localizadas na região Noroeste de Portugal, em que os dados foram observados num período de dois anos (de Janeiro de 2015 até Dezembro de 2016). As metodologias desenvolvidas contribuíram para o aumento da eficiência de gestão, em particular da eficiência energética e da sustentabilidade dos sistemas de tratamento e exploração de águas residuais.

Palavras e frases chave: ETAR, variáveis ambientais, físico-químicas, custos, sazonalidade, modelos lineares.

1. INTRODUÇÃO

As Estações de Tratamento de Águas Residuais (ETAR) têm um papel fulcral na conservação da Saúde Pública e na preservação da qualidade do meio receptor, minimizando os problemas de poluição da água causados por descarga de águas residuais não tratadas em meios hídricos. Com o objectivo de se obter um efluente com valores de qualidade compatíveis com os limites de descarga estabelecidos pelos normativos, são seleccionados o tipo de processo e as operações constituintes de uma ETAR, em particular o tipo de tratamento biológico. O principal processo biológico utilizado o tratamento por lamas activadas que consiste na produção de uma massa activada de microrganismos capazes de degradar a matéria orgânica por via aeróbia (processos que ocorrem na presença de oxigénio).

O tratamento de águas residuais é um processo de uso intensivo de recursos, principalmente energia. O consumo de energia representa uma parte significativa dos custos operacionais de uma estação de tratamento de águas residuais. Estudos recentes (Venkatesh e Brattebo, 2011, Elías-Maxil *et al.*, 2014) demonstraram que, numa ETAR convencional, cerca de 25% a 40% dos custos operacionais são atribuíveis ao consumo de energia, sendo que 70% desses consumos ocorrem nos sistemas de arejamento do tratamento biológico. Outra questão importante é a presença de infiltração de águas pluviais superficiais nas redes de esgoto, que sobrecarregam os efluentes e diluem a carga orgânica, provocando comportamentos hidrodinâmicos anómalos (Panepinto *et al*., 2016).

Sendo absolutamente fundamental uma gestão apoiada na sustentabilidade e na estabilidade dos processos de tratamento de águas residuais, importa procurar obter da modelação estatística o alicerce necessário à fundamentação de determinados fenómenos verificados na exploração corrente dos sistemas de saneamento, de modo a que seja possível entender e prever comportamentos. O uso da modelação estatística e análises complementares é seguramente um contributo positivo no controlo do desempenho das ETAR, nomeadamente na rentabilização dos custos de exploração.

2. METODOLOGIA

Este trabalho envolveu o estudo de nove ETAR localizadas na região Noroeste de Portugal (cinco localizadas em regiões rurais e quatro em regiões urbanas). O estudo centrou-se na análise de quatro variáveis físico-químicas das águas residuais (que são tributárias das ETAR) e duas variáveis hidro-meteorológicas (precipitação pluviométrica (mm), número de dias (por mês) de precipitação pluviométrica).

As variáveis físico-químicas foram SSV - Sólidos Suspensos Voláteis (mg/l), SST - Sólidos Suspensos Totais (mg/l), CQO - Carência Química de Oxigénio (mg O_2/l) e CBO5 - Carência Bioquímica de Oxigénio em 5 dias (mg O_2/l) (Barros *et al.*, 1995).

O estudo envolveu também a análise dos volumes de afluentes de águas residuais para as ETAR (m^3) e duas variáveis económicas (variáveis de custo associadas aos tratamentos): consumo de energia KW/h (euros), lamas e subprodutos (euros).

Inicialmente, foi realizada uma análise exploratória de dados com o objectivo principal de identificar e caracterizar a gestão de processos das ETAR, particularmente os processos biológicos utilizados no tratamento de efluentes (sistemas de lamas activadas). Métodos baseados em procedimentos estatísticos foram desenvolvidos para quantificar e caracterizar a variabilidade das variáveis de qualidade tanto do esgoto bruto quanto do efluente primário que

abastece os reactores de aeração da maioria das Estações de Tratamento de Águas Residuais (ETAR).

A principal discussão centrou-se na formulação de modelos estatísticos (Costa e Gonçalves, 2012) para quantificar e caracterizar padrões de interacção entre os afluentes das águas residuais às ETAR, as variáveis hidro-meteorológicas (como a pluviosidade), as variáveis físico-químicas e as variáveis de custo associados aos tratamentos. Em particular, a correlação entre as variáveis físico-químicas de controlo da eficiência dos processos, e destes com variáveis meteorológicas (pluviosidade), associada à definição criteriosa de funções-objectivo, foi essencial para a produção de novo conhecimento que suporte a minimização dos recursos utilizados nos sistemas de tratamento de águas.

Modelos de calibração e modelos lineares (Gonçalves e Alpuim, 2011) foram estabelecidos a partir de uma análise exploratória de dados visando estimar e prever os procedimentos de monitorização dinâmicos envolvidos nesses processos. Esses modelos integraram os comportamentos sazonais ao longo do ano, que têm um enorme impacto nas variações dos processos, considerando os dados observados nas estações seca e chuvosa. O processo de modelação considerou dois períodos hidrológicos: a estação seca (Julho, Agosto e Setembro) e a estação chuvosa (Janeiro, Fevereiro, Março, Abril, Maio, Junho, Outubro, Novembro e Dezembro).

As metodologias desenvolvidas neste estudo incidiram essencialmente na exposição/justificação de certos fenómenos, e da sua relação de interdependência com diferentes variáveis, bem como os principais efeitos em matéria de custos de exploração. Este estudo relacionou variáveis físico-químicas, hidro-meteorológicas e económicas, tendo sido obtidos resultados que permitiram consubstanciar relações de dependência entre variáveis e efeitos que os técnicos que conduzem as instalações de tratamento apenas empiricamente conheciam, estando agora ao seu dispor esta análise, que lhes permite avaliar e prever comportamentos em diferentes subsistemas.

3. CONCLUSÕES

A modelação estatística é uma ferramenta de extrema relevância na optimização das operações dos sistemas de tratamentos de águas residuais, nomeadamente na eficiência do tratamento das águas residuais afluentes às ETAR (qualidade do efluente final), bem como na diminuição dos consumos energéticos, de subprodutos e respectivos custos associados (em particular a optimização da etapa de arejamento, a principal responsável pelo consumo energético de uma ETAR).

Com este estudo, espera-se que as metodologias estatísticas adoptadas tenham uma relevante contribuição para a eficiência da gestão, em particular, para a eficiência energética, sustentabilidade dos sistemas de tratamento e exploração de águas residuais, assim como para a protecção da saúde pública e dos ecossistemas aquáticos.

AGRADECIMENTOS

Marco Costa foi parcialmente financiado por fundos portugueses através do CIDMA (Centro de Investigação e Desenvolvimento em Matemática e Aplicações) e da FCT (Fundação para a Ciência e a Tecnologia), através do projecto UID/MAT/04106/2013. Este trabalho foi parcialmente financiado pelo Centro de Matemática da Universidade do Minho por Fundos

Nacionais através da FCT – Fundação para a Ciência e a Tecnologia, no âmbito do projecto PEstOE/MAT/UI0013/2017. Este trabalho foi financiado pelos fundos da FEDER através do Programa Operacional dos Fatores de Competitividade - COMPETE e pelos fundos nacionais através da FCT - Fundação para a Ciência e Tecnologia no âmbito do projecto POCI-01-0145-FEDER-007136.

Referências

[1] Gonçalves A.M., Alpuim T. (2011). Water quality monitoring using cluster analysis and linear models. *Environmetrics* 22(8), 933-945.

[2] Costa, M., Gonçalves, A.M., (2012). *Combining Statistical Methodologies in Water Quality Monitoring in a Hidrological Basin – Space and Time Approaches*. Water Quality Monitoring Assessment, ed. Kostas Voudouris and Dimitra Voutsa, 121-142. ISBN: 978-953-51-0486-5. InTech Published.

[3] Barros, M.C., Mesquita, M., Vieira, P., Silva, M.C. (1995). *Laboratórios de Análises de Águas e Resíduos*. Volume 11, Edição LNEC.

[4] Elías-Maxil JA., Peter van der Hoek J., Hofman J., Rietveld L. (2014). Energy in the urban water cycle: actions to reduce the total expenditure of fossil fuels with emphasis on heat reclamation from urban water. *Renew Sustain Energy Rev* 30, 808-820.

[5] Panepinto D., Fiore, S., Zappone, M., Genon, G., Meucci L. (2016). Evaluation of the energy of a large wastewater treatment plant in Italy. *Applied Energy* 161, 404-4

[6] Venkatesh G., Brattebo H. (2011). Energy consumption, costs and environmental impacts for urban cycle services: case study of Oslo (Norway). *Energy* 36, 792-800.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

MODIFICAÇÃO NO MODELO PROBIT PARA AVALIAÇÃO DA GERMINAÇÃO EM SEMENTES DE MILHO

Deoclecio Jardim Amorim[1], Rute Quelvia de Faria[2], Amanda Rithieli Pereira dos Santos[1] e Maria Márcia Pereira Sartori[1]

[1] Universidade Estadual Paulista "Júlio de Mesquita Filho"; Faculdade de Ciências Agronômicas. Botucatu/SP, Brasil

[2] Instituto Federal Goiano. Campus Urutaí/GO, Brasil

RESUMO

As sementes constituem o insumo agrícola mais importante e o principal veículo de propagação das plantas. Dessa forma, necessitam ser de alta qualidade. A qualidade das sementes pode ser avaliada pela germinação em função do tempo, sendo a variável avaliada binária, no entanto, os métodos utilizados para essa avaliação exigem uma distribuição normal (*Probit*), segundo o teorema do limite central uma aproximação mais adequada de uma distribuição binomial a uma normal deve conter uma correção conhecida como Correção de Continuidade. Portanto aplicou-se essa correção ao intervalo crescente delimitado pela análise de resíduos aos dados de germinação de três lotes de sementes de milho, após transformação *Probit*, obtendo-se resultados mais precisos. Conclui-se que o método *Probit* foi eficiente para avaliação da germinação quando os dados não apresentam resíduos autocorrelacionados.

Palavras e frases chave: Função de ligação, Regressão linear, Análise de resíduos, Germinação, *Zea mays* L.

1. INTRODUÇÃO

As sementes constituem o principal veículo de reprodução das plantas permitindo a ocupação de novos hábitats, e a forma de conduzir ao campo as características genéticas importantes de desempenho do cultivar. Sementes de alta qualidade contribuem decisivamente para o sucesso do estabelecimento do estande, oferecendo suporte para uma produção rentável [1].

Sementes para serem consideradas de qualidade necessitam apresentar altos percentuais de germinação, por exemplo, na cultura do milho (*Zea mays* L) os cultivares de híbridos simples necessitam de pelo menos 85% de germinação [2]. Nesse sentido, o entendimento do processo de germinação é primordial para a escolha de lotes de sementes de alta qualidade.

A qualidade das sementes pode ser avaliada pela germinação em função do tempo, sendo a variável avaliada binária, haja vista que a germinação pode ou não ocorrer dependendo da sua qualidade. Quando a variável resposta é binária os erros não tem distribuição normal assumindo dois possíveis valores possuindo variâncias heterogêneas e restrições na função resposta. Assim,

funções resposta que tem essas características são denominadas funções logísticas. Uma propriedade interessante é que a função logística pode ser linearizada, por transformações do tipo *Probit* [3].

A transformação realizada com a função de *Probit* permite a estimação do tempo para 50% das sementes germinarem (T_{50}) de uma amostra (lote). O T_{50} está intimamente ligado ao vigor de um lote, ou seja, sementes com menor tempo de germinação terão mais chances de sobreviver na lavoura [1]. Todavia, nem sempre os pressupostos exigidos por esta transformação e pela regressão são alcançados, o que em alguns casos impedem a obtenção do T_{50} [4], como também, permite a obtenção de estimativas com erro de até 40% [5]. Assim, o objetivo desse trabalho foi avaliar modificações que permitam a melhoria da avaliação da germinação pela função *Probit*.

2. METODOLOGIA

Para aplicação do método foram utilizadas três cultivares de milho de alto rendimento plantadas em grandes áreas em todo território brasileiro. Para isso, avaliou-se a germinação de 20 sementes onde foram semeadas em três folhas de papel tipo "germitest" umedecidas com água destilada equivalente a 2,5 vezes o peso do papel, sendo quatro repetições em placas de "petri" e mantidas em câmara de germinação, regulada para a temperatura de 20 a 30 ºC. A contagem das sementes germinadas foi efetuada em intervalos regulares até 204 horas adotando como critério de germinação a protrusão da raiz primária ≥ 2 mm, onde os resultados foram expressos em percentagem de germinação acumulada ao longo do tempo. As porcentagens acumuladas foram linearizadas pela função de ligação P*robit* dada por:

$$Probit = \Phi^{-1}(\mu_i) \quad (1)$$

Sendo $\boldsymbol{\Phi^{-1}}$ função de distribuição acumulativa inversa da distribuição normal e $\boldsymbol{\mu_i}$ a resposta média da i[a] linha.

Os dados após a linearização foram submetidos à análise de regressão simples e análise gráfica de resíduos. A análise gráfica foi utilizada para demarcar em quais pontos ocorria à mudança de comportamento ou dependência dos erros, definindo o intervalo crescente em que os mesmos ocorriam (Correção de Continuidade). Após definido o intervalo crescente, o mesmo foi corrigido para variar de -5 a 5 probits.

Os dados experimentais, sem nenhuma correção e corrigidos, foram submetidos à análise de regressão simples permitindo o cálculo do T_{50} utilizou-se a seguinte formula:

$$T_{50} = \frac{\beta_n}{\alpha_n} \quad (2)$$

Sendo $\boldsymbol{\beta_n}$ o intercepto das n equações lineares e $\boldsymbol{\alpha_n}$ o ângulo de inclinação das n equações lineares.

Para verificar a qualidade do ajuste das regressões lineares aos dados sem nenhuma correção e corrigidos foi utilizado o coeficiente de determinação ajustado (R^2 aj.), o teste de Durbin-Watson para detectar a presença de autocorrelação e o valor de T_{50}, sendo este comparado com intervalo experimental esperado (ocorrência de 50% germinação das sementes viáveis). A análise dos dados foi realizada pelo PROC REG do software SAS 9.0 [6]

3. RESULTADOS E DISCURSÃO

Os dados de germinação quando transformados em probitos e submetidos à análise de regressão simples e análise de resíduo demostraram pontos tendenciosos, apontando indícios de dependência dos resíduos (Figura 1).

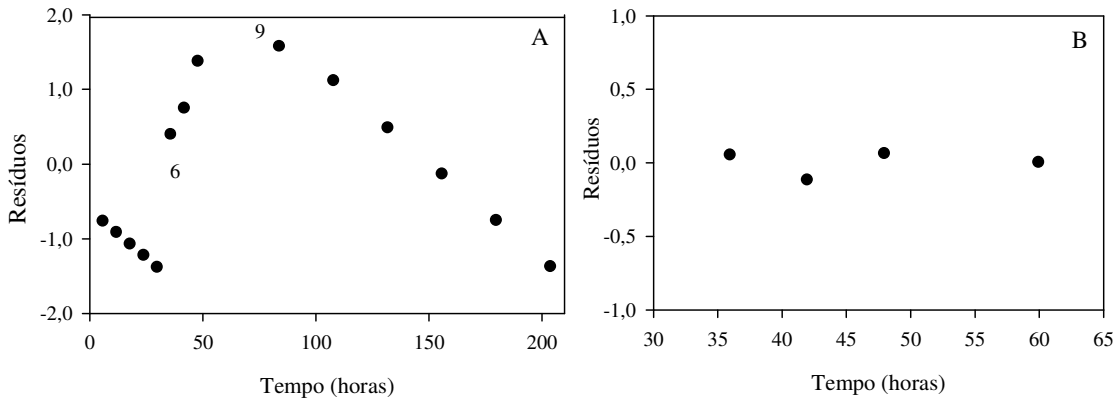


Figura 1: Resíduos versus tempo de avaliação dos dados sem correção (A) e com correção de continuidade (B).

Para um ajuste adequado de um modelo é necessário que o procedimento estatístico tenha suas exigências cumpridas. No caso de modelos lineares um dos pressupostos é a independência dos resíduos. Quando os resíduos não são independentes eles podem ter autocorrelação positiva ou negativa, ambas causam prejuízos nas estimavas dos parâmetros, diminuindo ou amentando a estimativa do erro da variância, respectivamente [7].

A Figura 1 é a representação dos resíduos versus tempo de avaliação dos dados da cultivar 01, essa mesma tendência foi observada para as cultivares 02 e 03. O comportamento dos resíduos, dos dados sem correção, mostra que o ajuste não é indicado, ou seja, devemos escolher uma função de ordem superior ou delimitar o intervalo de interesse que obedeça a distribuição normal (Figura 1A). Considerando o intervalo de interesse, e utilizando a correção de continuidade definida pela aproximação de uma distribuição binomial a uma normal conforme o Teorema do limite central, o intervalo de pontos de seis e nove foi selecionado, observando que esse intervalo da função é crescente. A Correção de continuidade é importante, pois permite que a função de ligação *Probit* seja usada e forneça estimativas mais precisas [3].

Observando o primeiro parâmetro, R^2 ajustado. nota-se que as equações lineares ajustadas aos dados sem a correção de continuidade obtiveram coeficiente de determinação ajustado menores que aqueles com correção. Menores coeficientes de determinação ajustados indicam baixa relação entre as variáveis, ao passo que maiores coeficientes são preferíveis indicando melhor qualidade de ajuste (Tabela 1) [8].

Lote	Coeficientes	Estimação	R^2 aj.	Durbin-Watson	T_{50}	Intervalo experimental
01 Sem correção	α	0,04103	0,86	0,439 **	86,8	48 a 60 horas
	β	-3,56136				
01 Corrigido	α	0,10454	0,97	2,02 ns	53,2	
	β	-5,56096				
02 Sem correção	α	0,0399	0,65	0,323 **	56,4	36 a 42 horas
	β	-2,25198				
02 Corrigido	α	0,16853	0,96	2,202 ns	39,9	
	β	-6,7233				
03 Sem correção	α	0,03781	0,79	0,501 **	56,6	42 a 48 horas
	β	-2,14166				
03 Corrigido	α	0,0881	0,97	2,24 ns	45,4	
	β	-4,00029				

Tabela 1: Parâmetros utilizados para verificar a qualidade de ajuste, significativo a 1% (**) e não significativo (ns).

O teste de Durbin-Watson foi significativo a 1% de probabilidade para as equações lineares ajustadas aos probitos da germinação sem nenhuma correção demostrando que todas possuíam resíduos autocorrelacionados positivamente tornando a estimativa do erro da variância muito pequena causando prejuízos na estimação dos coeficientes da regressão linear. Assim, os valores de T_{50} foram superestimados quando comparados aos resultados experimentais.

O teste de Durbin-Watson foi não significativo quando as equações lineares ajustadas aos probitos da germinação foram corrigidos, indicando que os resíduos são independentes (Figura 1B), não prejudicando a estimativa dos coeficientes de regressão. Fato comprovado pelos valores de T_{50} encontrados dentro dos intervalos experimentais esperados.

4. CONCLUSÕES

1. A correção de continuidade deve ser aplicada para toda avaliação de germinação que após a linearização por *Probit* os dados apresentem resíduos autocorrelacionados.

3. O método *Probit* aplicado aos dados corrigidos foi eficiente para a determinação do T_{50} nos três cultivares testados, apresentando resultados próximos aos dados experimentais. Assim este método, apresenta vantagens em relação aos tradicionais métodos utilizados para o cálculo desse importante parâmetro no que tange a qualidade de sementes, haja vista ser de fácil aplicabilidade.

Referências

[1] Marcos Filho, J. (2005). Fisiologia de sementes de plantas cultivadas. Piracicaba: Fealq, São Paulo.

[2] Brasil, (2013). Instrução Normativa n° 45, de 17 de setembro de 2013.

[3] Freitas, L. R. et al. (2013). Comparação das funções de ligação Logit e Probit em regressão binária considerando diferentes tamanhos amostrais Enciclopédia Biosfera, Centro Científico Conhecer - Goiânia, v.9, n.17; p. 29-36.

[4] Hill, H. J. et al. (2007). Primed Lettuce Seeds Exhibit Increased Sensitivity to Moisture Content During Controlled Deterioration. HortScience, v.42, p.1436-1439.

[5] Sartori, M. M. P. et al. (2017). Desenvolvimento de um software para avaliação da germinação e longevidade de sementes. 2017. IN: XX Congresso Brasileiro de Sementes (ABRATES), Anais... Foz do Iguaçú/ SP.

[6] Statistical Analyses System. (2003). Version Release 9.0 for Windows. Cary: (CD-ROM).

[7] Sas Institute et al. (2008). SAS/STAT 9.1 User's Guide the Reg Procedure:(Book Excerpt). SAS Institute.

[8] Araújo, E. J. G. et al. (2012). Relação hipsométrica para candeia (*Eremanthus erythropappus*) com diferentes espaçamentos de plantio em Minas Gerais, Brasil. Pesquisa Florestal Brasileira, v. 32, n. 71, p. 257–268.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

PREDIÇÃO DINÂMICA DA SOBREVIVÊNCIA A LONGO PRAZO EM DOENTES COM CANCRO DA MAMA

Sofia Azevedo[1], Susana Esteves[2] e Lisete Sousa [1,3]

[1] Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal.
[2] Instituto Português de Oncologia de Lisboa, Francisco Gentil, E.P.E, Portugal.
[3] Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal.

RESUMO

Em doentes com cancro de mama o risco de recaída e a sobrevivência esperada são habitualmente estimados no momento do diagnóstico com base em fatores clínicos e anatomo-patológicos. Embora a estratificação do risco no momento do diagnóstico assuma uma importância central na decisão terapêutica inicial, as estimativas de sobrevivência baseadas nas curvas de sobrevivência tradicionais no momento do diagnóstico poderão não facultar informação rigorosa quanto ao prognóstico a longo prazo. Em vários tipos de cancro já foi demonstrado que o risco de morte ou recaída é maior nos primeiros anos após o diagnóstico, tendendo a diminuir com o passar do tempo, ou seja, a probabilidade de sobrevivência por um período adicional de tempo aumenta à medida que os doentes vivem mais tempo para além do diagnóstico inicial de cancro. Nestas circunstâncias, e tendo em conta que o prognóstico melhora com o passar do tempo, as estimativas iniciais tornam-se menos relevantes à medida que o tempo desde o diagnóstico aumenta. Assim, os métodos de predição dinâmica são os mais adequados para prever a evolução da doença condicional face à situação atual do paciente. Neste estudo considera-se uma coorte de mulheres com cancro de mama em estádio inicial que receberam tratamento de acordo com as práticas padrão atuais, tendo como principais objetivos: 1) caracterizar a sobrevivência global, sobrevivência livre de doença e a probabilidade de recaída condicional ao tempo vivido sem doença (0-5 anos) e 2) avaliar a signficância a longo prazo de fatores de prognóstico relevantes inicialmente e avaliar como os seus efeitos variam com o tempo. A sobrevivência condicional vai ser definida como a probabilidade de continuar vivo ou livre de doença por mais 2 e 5 anos, sabendo que a paciente está livre de doença há 0, 1, 2, 3, 4 e 5 anos. Estas análises irão ser implementadas em grupos definidos por estádio de doença (I, II e III), grau do tumor (baixo, intermédio ou alto), estado dos nódulos (positivo e negativo) e subgrupos biológicos: luminal A (ER $\geq$ 60%, PgR $\geq$ 20%, HER-2 negativo), luminal B (ER<60%, PgR<20%, HER-2 negativo), HER-2 positivo e triplo negativo (ER, PgR e HER-2 negativo). A análise multivariável vai ser inicialmente implementada através de um modelo de regressão de Cox, considerando variáveis clínicas e demográficas: idade, estado dos nódulos,

ER, PgR, HER-2, estádio do tumor, grau e tratamento inicial. Estas variáveis foram selecionadas inicialmente com base na sua relevância clínica e na sua significância no prognóstico inicial, de acordo com a literatura existente. Os resíduos de Schoenfeld serão aplicados para avaliar o pressuposto de proporcionalidade, para cada variável. Avaliar-se-á também de que forma o efeito das covariáveis se altera com o tempo, através do uso de novas medidas de predição obtidas no âmbito de desenvolvimentos teóricos recentes para a predição dinâmica de um indivíduo permanecer vivo em determinados instantes temporais posteriores.

Palavras e frases chave: Análise de Sobrevivência, Predição Dinâmica, Cancro da Mama.

AGRADECIMENTOS

Este trabalho é financiado por Fundos Nacionais através da FCT - Fundação para a Ciência e a Tecnologia no âmbito do projecto UID/MAT/00006/2013.

Referências

[1] Merrill, RM. (2017). Conditional relative survival among female breast cancer patients in the United States. *Breast Journal*, 00:1-3.

[2] Van Houwelingen, HC., Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis*. Ed. Boca Raton: CRC Press.

[3] Buzdar, AU. (2006). Aromatase inhibitors: changing the face of endocrine therapy for breast cancer. *Breast Disease*, 24, 107-117.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

ANÁLISE ESTATÍSTICA DAS TEMPERATURAS MENSAIS DO AR NO PORTO – MODELAÇÃO DE ESPAÇO DE ESTADOS NO PERÍODO DE 1888 A 2001

Marco Costa[1,2] e Magda Monteiro[1,2]

[1]ESTGA – Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro
[2]CIDMA – Centro de Investigação e Desenvolvimento em Matemática e Aplicações, Universidade de Aveiro

RESUMO

Nas últimas décadas, o mundo tem sido confrontado com as consequências do aquecimento global. No entanto, esse fenómeno global não se reflete igualmente em todas as partes do globo. Este trabalho analisa a série temporal de longo-prazo das temperaturas médias mensais do ar na cidade do Porto, Portugal. Neste trabalho propomos um modelo de espaço de estados com estado periódico cujos resultados indicam que existem diferentes taxas de aumento da temperatura estimando-se um aumento médio anual da temperatura de $2,17°C$, por século.

Palavras e frases chave: alterações climáticas, filtro de Kalman, modelos de espaço de estados.

1. INTRODUÇÃO

O aumento da temperatura global tem sido uma preocupação crescente de várias autoridades. Segundo o Painel Intergovernamental sobre Mudanças Climáticas, as emissões mundiais de gases de efeito estufa continuam a aumentar, sendo que este aumento excederá em muito a meta limite de dois graus Celsius acordada pelos países no âmbito do Acordo de Paris.
Em particular, tem havido mudanças climáticas especialmente relevantes na Península Ibérica. De facto, um grande aumento nas temperaturas foi observado nos últimos 50 anos na Península Ibérica e, nos últimos 30 anos, o aquecimento ocorreu principalmente no verão ([5]). Assim, no contexto europeu e, em particular, na Península Ibérica, a análise de séries temporais locais tem um interesse especial, a fim de monitorizar o aumento da temperatura.
Os dados da temperatura do ar podem ser diários, mensais ou anuais, dependendo da natureza da escala, do tema a analisar e do histórico temporal disponível. Contudo, muita investigação tem sido desenvolvida com base em dados mensais de temperatura ([3, 2, 1].
Neste estudo, abordamos o problema da modelação de séries temporais mensais de temperatura através da formulação de um modelo de espaço de estados, associado a uma versão adequada do filtro de Kalman, que incorpora efeitos fixos e componentes estocásticas considerando uma estrutura periódica.

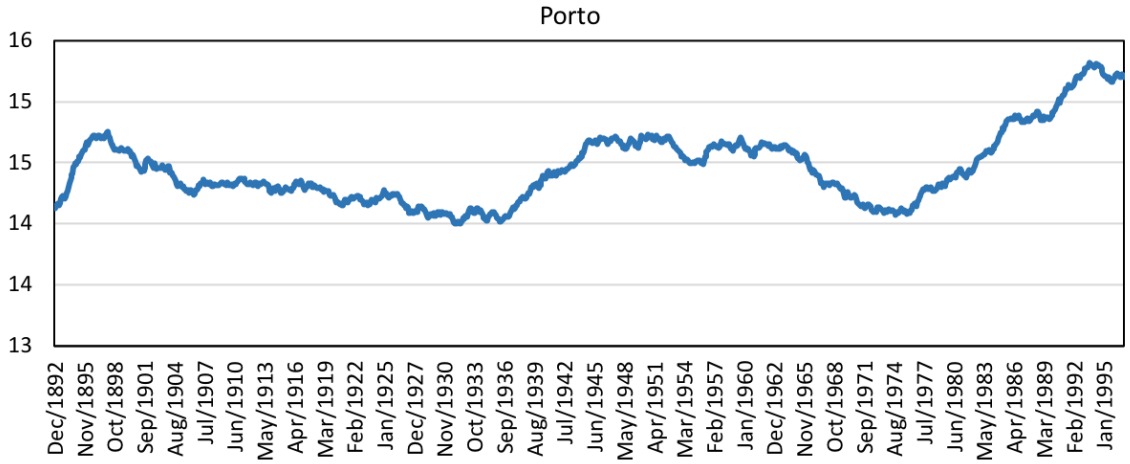


Figura 1: Média móvel centrada de 10 anos da temperatura média mensal do ar no Porto.

O objetivo deste trabalho é analisar a série temporal da temperatura mensal do ar na cidade do Porto. [6] estudou estes dados, a fim de detectar e corrigir mudanças não climáticas. Este conjunto de dados está disponível em [7].

2. DESCRIÇÃO DOS DADOS

O conjunto de dados original foi medido pelo Instituto Geológico do Observatório da Serra do Pilar da Universidade do Porto (IGUP), Porto, a partir de 1888 até 2001, compreendendo 114 anos (1368 observações).
Este trabalho incide sobre a série temporal da temperatura média obtida a partir da média mensal das semi-amplitudes térmicas diárias, $Y = (T_{min} + T_{max})/2$, uma vez que esta é uma variável climática muito considerada na investigação em alterações climáticas.
Como a série temporal é bastante longa, na Fig. 1 apresenta a média móvel de 10 anos da série temporal para facilitar uma inspeção visual dO comportamento geral. Nesta figura é claro que o aumento da temperatura difere em diferentes períodos de tempo.

3. MODELO EM ESPAÇO DE ESTADOS COM ESTADO PERIÓDICO

O modelo em espaço de estado periódico (MEEP) considera a variável observável Y, a temperatura média mensal do ar durante 114 anos, tendo cada ano 12 *estações* (meses). Assim, denotamos $Y_t \equiv Y_{s,n}$ com $t = 1, 2, ..., 1368$, $n = 1, 2, ..., 114$ e $s = 1, 2, ..., 12$, onde n é o ano associado com o mês t e s é o respetivo mês. Com essa notação, quando t corresponde a um janeiro do ano n, o mês anterior, que é o mês 12 do ano $n-1$, pode ser indicado, por simplicidade, como o mês 0 do ano n, ou seja, $Y_{0,n} \equiv Y_{S,n-1}$.
O modelo é definido através das seguintes equações:

$$Y_{s,n} = \begin{bmatrix} 1 & S(n-1)+s \end{bmatrix} \begin{bmatrix} a_{s,n} \\ X_{s,n} \end{bmatrix} + D_{s,n}\beta + e_{s,n} \tag{1}$$

$$\begin{bmatrix} a_{s,n} \\ X_{s,n} - \mu_s \end{bmatrix} = \begin{bmatrix} \phi_a & 0 \\ 0 & \phi_s \end{bmatrix} \begin{bmatrix} a_{s-1,n} \\ X_{s-1,n} - \mu_{s-1} \end{bmatrix} + \begin{bmatrix} \omega_{s,n} \\ \varepsilon_{s,n} \end{bmatrix}. \tag{2}$$

A matriz linha $\mathbf{H}_{s,n} = \begin{bmatrix} 1 & S(n-1)+s \end{bmatrix}$ representa a matriz de planeamento onde, em particular $S(n-1)+s$ representa o tempo. O vetor aleatório $\mathbf{X}_{s,n} = \begin{bmatrix} a_{s,n} & X_{s,n} \end{bmatrix}'$ tem uma estrutura de um Vetor Autorregressivo Periódico (PVAR) de ordem 1 que inclui: um processo autorregresivo não-periódico, $a_{s,n} \equiv a_t$, que representa a correlação mês-a-mês; e

um processo periódico autorregressivo de ordem 1, PAR(1), $\{X_{s,n}\}$ que representa os declives estocásticos.
O modelo incorpora efeitos sazonais fixos representados pelo vetor $\beta = [\beta_1\ \beta_2\ ...\ \beta_{12}]'$. A matriz de planeamento $D_{s,n}$ contém 0's and 1's, através de funções indicatrizes para associar β_s ao respetivo mês da variável $Y_{s,n}$. O erro de observação $e_{s,n}$ é um ruído branco gaussiano com variância $\text{Var}(e_{s,n}) = \sigma_e^2$.
Na equação de estado, Eq. 2, o *estado* $\mathbf{X}_{s,n}$ segue um PVAR(1) com média $\mu_{\mathbf{X}_{s,n}} = [\ 0 \quad \mu_s\ ]'$, onde μ_s é a média do declive do mês s; Φ_s é a matriz de parâmetros autorregressivos $\Phi_s = \text{diag}\{\phi_a, \phi_s\}$, onde ϕ_a é o coeficiente autorregressivo do processo AR(1), $\{a_{s,n}\}$, e ϕ_s é o coeficiente autorregressivo associado ao declive do mês s. O vetor de erros $\zeta_{s,n} = [\ \omega_{s,n} \quad \varepsilon_{s,n}\ ]'$ segue uma distribuição normal multivariada com matriz de covariâncias $\Sigma_{\zeta_{s,n}} = \text{diag}\{\sigma_\omega^2, \sigma_{\varepsilon,s}^2\}$, tal que,

$$\text{Cov}(\varepsilon_{s,n}, \varepsilon_{s-i,n}) = \begin{cases} \sigma_{\varepsilon,s}^2, & i = 0 \\ 0, & i \neq 0 \text{ for } i = 1, 2, ..., 12 \end{cases}$$

e os processos $\{\omega_{s,n}\}$ e $\{\varepsilon_{s,m}\}$ são não-correlacionados, tais que $\text{E}(\omega_{s,n}\varepsilon_{r,m}) = 0$, $\forall s, r, n, m$.
Esta abordagem permite a incorporação de algumas características que tornam o modelo versátil. No contexto da modelação da temperatura do ar, a abordagem de efeitos mistos associada à sazonalidade intra-anual é uma maneira simples de modelar a sazonalidade que naturalmente existe neste tipo de dados. O modelo de espaço de estados tem na sua estrutura um processo latente, o estado, que não é observável e precisa ser estimado. O procedimento mais comum para fazer esta previsão é o algoritmo do filtro de Kalman. Este algoritmo calcula, a cada momento, o estimador ótimo do vetor de estado baseado na informação disponível até ao instante t e o seu sucesso está no facto de que é um procedimento de estimação em tempo-real. Quando os erros e o estado inicial são gaussianos, os preditores do filtro de Kalman são os melhores estimadores não-viesados, no sentido do erro quadrático médio mínimo.
No entanto, as propriedades ótimas somente podem ser garantidas quando todos os parâmetros do modelo Θ forem conhecidos ([8]). Quando parâmetros do modelo de espaço de estados são estimados, a incerteza associada aos os estimadores de filtro de Kalman são subestimados e alguns procedimentos podem ser implementados ([4]).

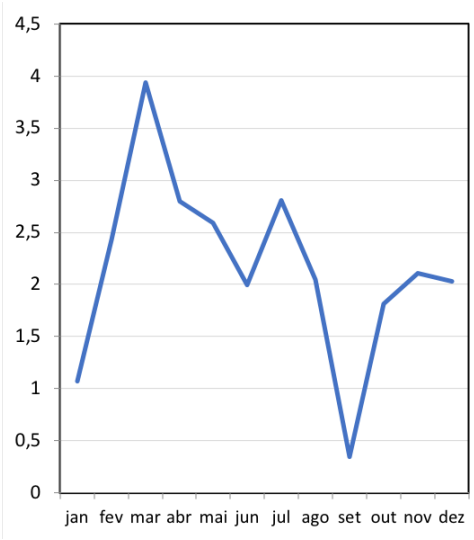
4. ALGUNS RESULTADOS

Antes da discussão e interpretação dos resultados, foram realizados vários procedimentos para validar o modelo e avaliar os respetivos pressupostos. Numa análise global, todo o modelo ajustou-se bem aos dados, uma vez que todos os pressupostos são verificados e também tem associado um elevado coeficiente de determinação ($R^2 = 0,9189$).
As inovações (erros de previsão a 1 passo), $\eta_{s,n}$, foram consideradas como tendo uma distribuição gaussiana condicional $\eta_{s,n} = Y_{s,n} - \widehat{Y}_{s|s-1,n} \sim N(0, \omega_{s,n})$.
A normalidade da série das inovações foi testada considerando-se o teste de Kolmogorov-Smirnov (K-S) e o teste Jarque-Bera (JB). Em ambos os testes, a normalidade não foi rejeitada, considerando os usuais 5% para a significância, sendo ambos os valores-p superiores a $0,20$. Além disso, o histograma e o gráfico QQ com envelopes de confiança de 95 % das inovações padronizadas indicam que a distribuição empírica é concordante com a curva de normal. A série das inovações não apresenta correlação temporal, uma vez que a função de autocorrelação e autocorrelação parcial empíricas, FAC e FACP indicaram que as inovações são compatíveis com um processo de ruído branco. Os principais resultados obtidos estão apresentados na Fig. 2.

5. CONCLUSÕES

mês	s	$\mu_s \cdot 100 \cdot 12$	ϕ_s	σ^2_ϵ
jan	8,426	1,076	0,428	0,330
fev	8,184	2,427	3,029	0,909
mar	8,803	3,939	1,265	0,749
abr	11,039	2,801	0,907	0,661
mai	13,323	2,594	1,165	0,633
jun	16,692	1,998	0,755	0,394
jul	17,766	2,810	1,326	0,475
ago	18,370	2,047	0,712	0,295
set	18,580	0,343	0,227	0,658
out	14,604	1,811	2,398	1,029
nov	10,666	2,109	1,231	0,720
dez	8,429	2,031	1,087	0,871

$\prod_{s=1}^{12} \phi_s = 0,9$



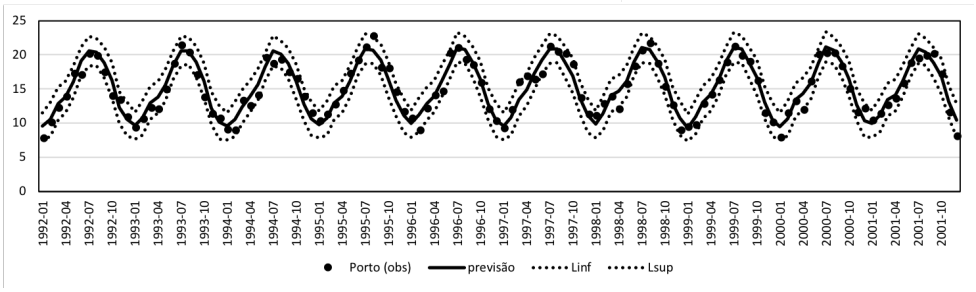


Figura 2: Estimativas dos parâmetros do modelo e previsões a 1 passo de 1991 a 2001.

O modelo proposto foi ajustado à série temporal de longo-prazo da temperatura média mensal do ar do Porto, verificando-se os respetivos pressupostos. Os principais resultados mostraram que se estima que no Porto o aumento da temperatura média mensal foi 2.1655° C, por século, no período analisado. Contudo, o aumento foi diferenciado para cada mês do ano.

AGRADECIMENTOS

Os autores foram parcialmente financiados por fundos portugueses através do CIDMA e da FCT (Fundação para a Ciência e a Tecnologia), através do projeto UID/MAT/04106/2013.

Referências

[1] Abbasnia, M., Toros, H. (2016). Future changes in maximum temperature using the statistical downscaling model (SDSM) at selected stations of Iran. *Modeling Earth Systems and Environment*, 2(68), doi:10.1007/s40808-016-0112-z

[2] Alpuim, T., El-Shaarawi, A. (2009). Modeling monthly temperature data in Lisbon and Prague. *Environmetrics*, 20: 835–852.

[3] Bengtsson, T., Cavanaugh, J.E. (2008). State-space discrimination and clustering of atmospheric time series data based on Kullback information measures. *Environmetrics*, 19, 103–121.

[4] Costa, M. & Monteiro, M. (2016). Bias-correction of Kalman filter estimators associated to a linear state space model with estimated parameters. *Journal of Statistical Planning and Inference*, *176*, 22–32.

[5] European Environment Agency (2017). Global and European temperatures, disponível no sítio web em `https://www.eea.europa.eu/data-and-maps/indicators/global-and-european-temperature-4/assessment`

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

MODELOS LONGITUDINAIS PARA MOMENTOS DE INOVAÇÃO EM PSICOTERAPIA

Gina da Silva Voss[1] e Inês Pereira Silva Cunha de Sousa[2]

[1]Universidade do Minho
[2]Universidade do Minho

RESUMO

O paciente diagnosticado com depressão/ansiedade, submetido ao tratamento com sessões de terapia, passa por mudanças psicológicas e comportamentais, as quais foram denominadas momentos de inovação (MIs). A motivação desse estudo foi estudar como esses momentos de inovação contribuem para a melhora do paciente. Para tal, usamos os modelos longitudinais, onde se leva em consideração a correlação existente entre as medidas de um mesmo indivíduo. O objetivo principal é a aplicação de modelos longitudinais para a análise dos momentos de inovação em sessões de psicoterapia, e suas relações com outras variáveis de interesse (sintomatologia, aliança paciente - terapeuta e retornos ao problema). Os dados do presente estudo, são provenientes de um serviço de psicologia da cidade de Braga - PT, na qual foram observados 63 pacientes diagnosticados com depressão/ansiedade, ao longo de 20 sessões de terapia, e, em cada sessão, recolheram-se informações sobre a sintomatologia ($OQ10$), a aliança que se forma entre paciente-terapeuta (WAI), os momentos de inovação (MIs) e os retornos ao problema ($RPMs$). As sessões de terapia foram realizadas pelos métodos *cognitivo* ou *narrativo*, onde analisamos a sua influência sobre as variáveis psicoterapêuticas. Após as análises, concluímos que as variáveis psicoterapêuticas estão a influenciar umas às outras, com exceção dos $RPMs$ como preditores da sintomatologia e da aliança paciente-terapeuta. Os momentos de inovação foram divididos em dois grupos (*high* e *low*) e analisamos a sua influência na sintomatologia $OQ10$. Os $MI.high$ são significativos para explicar a sintomatologia, enquanto que os $MI.low$ não o são.

Palavras e frases chave: momentos de inovação; modelos longitudinais; psicoterapia; dados correlacionados.

1. INTRODUÇÃO

Depressão é uma doença grave, de carácter psicológico, que afeta diferentes tipos de pessoas, sendo que, cada indivíduo é afetado de maneira única. Os danos causados geram impacto direto e profundo na vida do paciente e seus familiares. A Organização Mundial de Saúde classifica a depressão como o maior contribuinte da incapacidade para a atividade produtiva,

registando, em 2015, 7,5% de todos os anos vividos com incapacidade [1]. Em 2013, com o 1^o Estudo Epidemiológico Nacional de Saúde Mental (parte integrante do World Mental Health Survey Initiative, da Organização Mundial de Saúde e da Harvard University), conseguiu-se olhar a depressão em números. Portugal tem a 4^o posição dentre os 34 países participantes do estudo, estando atrás do Brasil, EUA e Irlanda do Norte [2].
Durante as sessões de terapia, mudanças pessoais são esperadas na narrativa do paciente, pois a psicoterapia atua num nível mais profundo, onde as auto-narrativas problemáticas acabam por transformarem-se em auto-narrativas com mais esperança e auto-estima. Essas mudanças foram codificadas e denominadas momentos de inovação [3]. Foram classificadas em 5 tipos diferentes: ação (A), protesto (P), reflexão (R), re-conceptualização (RC), e mudança de performance (PC), sendo que os tipos ação (A), protesto (P) e reflexão (R) têm dois níveis, 1 e 2. Atualmente, os MIs estão a ser divididos em dois grupos chamados de $high$ (alto) e low (baixo). Os $MI.high$ são compostos pelos seguintes tipos de MIs: ação (A1), reflexão (R1), protesto (P1), re-conceptualização (RC), e mudança de performance (PC); enquanto os $MI.low$ são formados pelos MIs: ação (A2), reflexão (R2) e protesto (P2).

Os dados do presente estudo, são provenientes de um serviço de Psicologia da cidade de Braga - PT, onde foram observados 63 pacientes diagnosticados com depressão e/ou ansiedade. Várias variáveis foram observadas, tais como: a sintomatologia ($OQ10$), ou 'nível' em que se encontra da doença, é uma escala medida através de um questionário ($Outcome\ Questionnaire$ 10.2) [4] preenchido no início de cada sessão de terapia; a aliança paciente-terapeuta (WAI), significa o quanto o paciente e terapeuta 'se entendem', sendo medida através de questionário ao final de cada sessão; e o *outcome*, que é medido através do questionário BDI - Beck Depression Inventory [5], realizado no início e no final do tratamento. A diferença dos valores dos questionários ($BDI_{fim} - BDI_{inicio}$) é que determina um *poor outcome* (onde o paciente não apresentou melhora nos sintomas da depressão), ou um *good outcome* (o paciente apresentou melhora nos sintomas da depressão), sendo que este último determina o sucesso da terapia.

Segundo Singer [6], estudos longitudinais têm interesse especial quando o objetivo é avaliar tanto mudanças globais, quanto individuais, ao longo do tempo. Os modelos longitudinais conseguem absorver, juntamente com as covariáveis, a estrutura de correlação existente entre medidas de um mesmo indivíduo. Estudos longitudinais permitem a distinção entre o grau de variação na variável resposta para um indivíduo ao longo do tempo e a variação entre diferentes indivíduos, aumentando a qualidade das interpretações [7]. Este trabalho tem como propósito a aplicação de modelos longitudinais para o estudo dos momentos de inovação em sessões de psicoterapia, e de suas relações com outras variáveis de interesse (sintomatologia, aliança paciente - terapeuta e retornos ao problema). As análises estatísticas foram realizadas com o auxílio do software R.

2. METODOLOGIA

Neste estudo estamos a investigar a relação existente entre as variáveis psicoterapêuticas, e a técnica estatística utilizada é a análise de regressão. Após uma análise descritiva dos dados, os dados foram analisados transversalmente (considerando independência das observações). Para as variáveis resposta contínuas ($OQ10$ e WAI), utilizamos modelos de regressão linear múltipla e para as que seguem uma distribuição Binomial (MI e RPM, apresentadas em proporção), utilizamos os modelos lineares generalizados, onde modelamos a razão de chances de ocorrência de MIs e $RPMs$. Os parâmetros do modelo ($\hat{\theta}$) são estimados pelo método da máxima verossimilhança.

Para a determinação do modelo que melhor se ajusta aos dados usamos o critério de Akaike (AIC) para as regressões lineares múltiplas e para os modelos lineares generalizados, utilizamos os resíduos *deviance*. Com o modelo definido para os dados transversais, realizamos os variogramas (que é uma representação de possível existência de correlação entre observações de um mesmo indivíduo) e a linha não tende a ser constante, logo, há evidências da existência de uma estrutura de correlação dentro do indivíduo e é necessária a aplicação de técnicas estatísticas que absorvam as mesmas, tal como os modelos longitudinais.

Os modelos longitudinais conseguem incorporar a correlação entre as respostas de um mesmo indivíduo ao longo do tempo e melhor ajustar-se ao dados. Os modelos longitudinais (ou modelos mistos) contêm efeitos fixos, efeitos aleatórios e um erro. Para as variáveis apresentadas em proporção (MIs e $RPMs$) usamos modelos lineares generalizados mistos. Nos modelos lineares generalizados mistos com função de ligação *logit* modelamos a razão de chances de ocorrência de MIs e $RPMs$ e não o valor total como anteriormente.

A escolha do modelo longitudinal que melhor se ajustou aos dados foi realizada através do menor valor de AIC (critério de Akaike), do maior valor da máxima verossimilhança, e através da análise gráfica do variograma teórico. Para os modelos lineares generalizados mistos, o modelo com menor função desvio é o de melhor ajustamento.

3. CONCLUSÕES

Com este estudo concluímos que, quando em estudo transversal, o tipo de terapia realizada (*cognitiva* ou *narrativa*) é significativa para explicar as variáveis MI e RPM, contudo, ao realizarmos o estudo longitudinal, com a correlação das medidas de um mesmo indivíduo sendo introduzidas no modelo, o tipo de tratamento realizado passa a não ter influência nas variáveis psicoterapêuticas estudadas ($OQ10$, WAI, MI, e RPM), logo, o paciente pode ser tratado em qualquer um dos dois métodos (*cognitivo* ou *narrativo*) que isto não irá influenciar nos resultados finais do tratamento terapêutico.

A sintomatologia do paciente ($OQ10$) é influenciada pela aliança (WAI) e pela ocorrência de MIs, sendo que os retornos ao problema não têm efeito sobre o score de $OQ10$. A aliança (WAI) é a variável com maior influência, quanto melhor a aliança formada com o terapeuta, melhor a sintomatologia do paciente será. O número de momentos de inovação também tem influência na sintomatologia, quanto maior o número de momentos de inovação, melhor a sintomatologia será (menor o score de $OQ10$). A aliança (WAI) é afetada pela sintomatologia e pelo número de ocorrências de MIs. A variável com maior influência na aliança é a sintomatologia, quanto melhor o paciente está da doença, maior será a aliança com o terapeuta. Os momentos de inovação também têm efeito significativo e quanto maior o número de MIs, melhor a aliança paciente - terapeuta (WAI).

Os momentos de inovação (MIs) são diretamente afetados pela sintomatologia ($OQ10$) e pela aliança com o terapeuta (WAI). A sintomatologia é a variável com maior influência, por isso, para cada score a mais de $OQ10$, a chance de ocorrerem momentos de inovação diminui 3,63% (quanto pior o paciente está da doença, menor a chance de ocorrerem MIs). Na aliança paciente - terapeuta, por sua vez, para cada score a mais na aliança, a chance de ocorrerem momentos de inovação aumenta 1,21% (quanto melhor a aliança, maior o número de MIs).

Ao analisarmos os dois grupos de momentos de inovação (*low* e *high*) em relação a influência deles sobre a sintomatologia, o grupo de $MI.low$ não é significativo ao explicar a melhora do

paciente, enquanto o grupo $MI.high$ é significativo (está a explicar a melhora do paciente), logo, os MIs dos tipos ação (A2), reflexão (R2), protesto (P2), re-conceptualização (RC), e mudança de performance (PC), têm maior influência na melhora do paciente.

Referências

[1] Direção Geral Da Saúde - DGS. (2016). *Portugal – Saúde Mental em Números 2015*. Direção-Geral da Saúde, Lisboa. ISSN: 2183-1505.

[2] Faculdade de Ciências Médicas. (2013). *Estudo Epidemiológico Nacional de Saúde Mental - 1º Relatório*. Universidade Nova de Lisboa, Lisboa.

[3] Gonçalves, M.M., Silva J.R. (2014). Momentos de inovação em psicoterapia: Das narrativas aos processos dialógicos. *Análise Psicológica* 32, 27–43.

[4] Lambert, M.J., Gregersen, A.T., Burlingame, G.M. (2004). The Outcome Questionnaire - 45. *Lawrence Erlbaum Associates Publishers*.

[5] Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry* 4, 561–571.

[6] Singer, J.M., Nobre, J.S., Rocha, F.M.M. (2017). *Análise de dados longitudinais: versão preliminar*. Universidade de São Paulo: São Paulo.

[7] Diggle, P.J., Heagerty, P., Liang, K., Zeger, S.L. (2002). *Analysis of Longitudinal Data*. Oxford University Press: Oxford.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

IMPLEMENTATION OF BOOTSTRAP METHODS FOR ACCURACY ASSESSMENT OF SPACE-TIME DATA MODELLING

Gustavo Soutinho[1], Raquel Menezes[2]

[1]University of Minho, Portugal.
[2]CBMA & Department of Mathematics and Applications, University of Minho, Portugal.

ABSTRACT

The large and small-scale variation of a spatio-temporal stochastic process can be separately estimated by using a two-stepwise approach. Firstly, a generalized linear model (GLM) is adopted, regarding the data distribution to approximate the trend and seasonality, by relaxing the assumption of non-correlated errors. This procedure provides point estimates of the regression parameters, specifying the large-scale variation. Secondly, the small-scale variation, imposed by the underlying dependence of the stationary residual, is estimated through a spatio-temporal variogram, such as the sum-metric model which accounts for the space-time interaction [1].

As a consequence of the assumption of independence of the residuals, the maximum likelihood estimates (MLEs) for the regression parameters usually give us over-optimistic standard errors [2] and, consequently, not enable us to identify whether an independent variable have or not a significant contribution into the response variable. Moreover, the common methods used to obtain the parameters' estimates of the spatio-temporal variogram (e.g. spatial, temporal and joint variances or ranges) do not provide the standard errors associated. To overcome these drawbacks, bootstrap approaches are integrated into the estimation of the large and small-scale variation components.

This work aims to compare parametric and non-parametric bootstrap methods and to propose alternatives to assess the accuracy of estimates associated to the parameters of the large and small-scale variations. The parametric bootstrap approach considers replicates, drawn from a multivariate normal distribution, with expectation defined by the trend model and covariance matrix obtained from the sum-metric variogram. This parametric method may be used to analyse the significance of the parameters in the large-scale variation component.

In this study the following independent variables were considered: location of the monitoring station (longitude and latitude); week reference (1 up to 212); type of site (background, industrial or traffic) and the type of environment (urban, suburban or rural). To model the seasonality of data, we used an harmonic regression, assuming a period equal to 52-week.

The non-parametric bootstrap approaches are based on *moving block* and *random block* bootstrap methods, two different ways of drawing the dependent time series observations, when

taking fixed data in space dimension. As in environmental sciences, typically, the spatial resolution, defined by the monitoring stations, is smaller than the time resolution, these sequential time blocks procedures may prove to be useful. For both methods, in this work, the overlapping of blocks in the time dimension is considered. The main idea of the *moving block* bootstrap consists of dividing the temporal data, $X_1, ..., X_T$ into blocks of consecutive observations of length l. Each new block slides δ time units, allowing for a total of $k+1$ blocks, defined as $(X_1, \ldots, X_l)$, $(X_{1+\delta}, \ldots, X_{l+\delta})$, $\ldots$, $(X_{1+k\delta}, \ldots, X_{l+k\delta})$, such that $l + k * \delta \leq T$. Under the *random block* bootstrap, replicates are defined by M blocks of consecutive observations with the same length l randomly selected from the start time between 1 and $T - l + 1$ [3].

The weekly average of the NO_2 between 1 January 2013 and 31 December 2016, from 50 monitoring stations located on Mainland of Portugal, are used to illustrate the differences among bootstrap methods. The NO_2 is a good marker for the exposure of the air quality and is among the main pollutants with significant impact on environmental and health problems.

Results reveal that the bootstrap approaches are particularly appropriate to distinguish non-significant independent variables in GLM, when adopting the two-stepwise approach presented in [1], as well as, they are useful to analyse the significance of estimates of variogram parameters. In the former case, we can conclude that, among all the covariates initially considered in GLM, just type of environment and type of site have significant influence on the response variable (NO_2).

The parametric bootstrap method is preferable when the data distribution is known but requires more computational costs, whereas a non-parametric method should be used when we wish to avoid distributional assumptions and is computationally faster. The random block bootstrap allows to improve the accuracy of the estimates due to the possibility of choosing a larger number of replicates.

Keywords and key sentences: Bootstrap methods, Spatio-temporal data, Geostatistics, Air pollution.

This research was financed by Portuguese Founds through FCT -"Fundação para a Ciência e Tecnologia".

References

[1] Menezes R., Piairo H., Garcia-Soidánand P., Sousa I. (2016). Spatial temporal modellization of the NO2 concentration data through geostatistical tools, *Journal of Statistical Methods & Applications*, issue1, 107-124.

[2] Monteiro A., Menezes R. and Silva M.E. (2017). Modelling spatio-temporal data with multiple seasonalities: the NO2 Portuguese case, *Spatial Statistics journal*, Vol. 22, Part 2, 371-387.

[3] Kreiss JP, Paparoditis E. Rejoinder (2011). Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society*, 40(4):393-5.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

ANÁLISE ESPACIAL DAS PARTÍCULAS PM10 NA ÁREA METROPOLITANA DE LISBOA

Paula Pereira[1] e Conceição Ribeiro[2]

[1]Escola Superior de Tecnologia de Setúbal, Instituto Politécnico de Setúbal/ Centro de Estatística e Aplicações, Universidade de Lisboa; paula.pereira@estsetubal.ips.pt
[2]Instituto Superior de Engenharia,Universidade do Algarve / CEPAC, Universidade do Algarve / Centro de Estatística e Aplicações, Universidade de Lisboa; cribeiro@ualg.pt

RESUMO

A poluição atmosférica é um dos problemas ambientais que provoca mais efeitos nocivos a curto e longo prazo, sendo, nas últimas décadas, uma preocupação internacional principalmente devido ao seu impacto na saúde humana e no meio ambiente. As partículas inaláveis têm sido associadas a vários efeitos na saúde pública, incluindo uma série de problemas respiratórios e cardiovasculares graves. Além disso, essas partículas e os seus componentes interagem de modo a contribuir para concentrações elevadas de outros poluentes do ar. As partículas inaláveis que irão ser analisadas neste trabalho referem-se às PM_{10}, partículas inaláveis com diâmetro inferior a 10 μm. A Organização Mundial de Saúde assim como a Comissão Europeia fornecem limiares para os níveis de poluição nocivos à saúde. Em relação às partículas PM_{10}, o limite diário estabelecido pela Comissão Europeia é de 50 $\mu g/m^3$ e não deve ser excedido mais de 35 dias por ano. Este trabalho centra-se na análise das partículas PM_{10} na Área Metropolitana de Lisboa, os dados disponíveis são diários e foram recolhidos durante o ano de 2015 em várias estações de medição. A Área Metropolitana de Lisboa é uma divisão administrativa de Portugal que inclui 18 municípios da Grande Lisboa e da Península de Setúbal. A área desta região é de 3 015,24 km^2, o que representa cerca de $3,4\%$ da área total de Portugal, no entanto esta área tem a maior concentração de população em Portugal. Em 2015 cerca de $27,2\%$ da população total de Portugal vivia na Área Metropolitana de Lisboa. O objetivo deste trabalho é analisar a estrutura espacial e temporal subjacente às particulas PM_{10} na Área Metropolitana de Lisboa e obter mapas de concentração. Com estes mapas pretende-se avaliar a exposição humana e avaliar a conformidade com as diretivas europeias e nacionais, em particular em locais onde não existem estações de medição. Para atingir este objetivo recorreu-se aos modelos bayesianos hierárquicos.

Palavras-chave: Poluição do ar; análise espacial; modelos Bayesianos hierárquicos.

AGRADECIMENTOS

O trabalho é financiado pela FCT - Fundação para a Ciência e a Tecnologia, Portugal, através do projecto UID/MAT/00006/2013.

Referências

[1] Cameletti M., Ignaccolo R. and Bande S. (2011). Comparing spatio–temporal models for particulate matter in Piemonte. *Environmetrics*, 22 (8), 985–996.

[2] Cameletti M., Lindgren F., Simpson D. and Rue H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, 97 (2), 109–131.

[3] Russo A., Trigo R.M., Martins H. and Mendes M.T., (2014). NO_2, PM10 and O_3 urban concentrations and its association with circulation weather types in Portugal. *Atmospheric Environment*, 89, 768–785.

[4] Sahu, S.K., Gelfand, A.E. and Holland, D.M. (2006) Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 61–86.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

MODELOS COMPETITIVOS PARA ANALISAR AS SÉRIES TEMPORAIS DA CONCENTRAÇÃO DO OXIGÉNIO DISSOLVIDO NO RIO VOUGA

Magda Monteiro[1,2] e Marco Costa[1,2]

[1] ESTGA - Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro
[2] CIDMA - Centro de Investigação e Desenvolvimento em Matemática e Aplicações, Universidade de Aveiro

RESUMO

Neste estudo pretende-se avaliar o desempenho de modelos competitivos para descrever os valores mensais do oxigénio dissolvido em estações de monitorização da qualidade da água no rio Vouga. O confronto é realizado entre um modelo de regressão com erros correlacionados e dois modelos de espaço de estados.

Palavras-chave: Modelo de espaço de estados, filtro de Kalman, regressão linear, oxigénio dissolvido.

1. INTRODUÇÃO

A avaliação da qualidade da água superficial é uma parte importante da monitorização ambiental, cuja avaliação pode predizer a qualidade da água e evitar problemas de saúde pública de diversos tipos e níveis. Um papel importante na monitorização da qualidade da água superficial tem sido atribuído à variável concentração de oxigénio dissolvido (OD) uma vez que este indicador resulta do impacto de um conjunto de fatores ambientais como a temperatura da água, temperatura e pressão do ar, a morfologia do leito do rio, o estado de limpeza da água, as fontes de poluição das águas superficiais.
A análise química de OD mede a quantidade de oxigénio gasoso dissolvido numa solução aquosa. Para que a água seja considerada de boa qualidade é necessário que esta possua níveis adequados de OD, um elemento necessário para todas as formas de vida. Quando os níveis de OD na água descem abaixo de 5 mg/l, a vida aquática é colocada sob pressão. Níveis de oxigénio que permanecem abaixo de 1-2 mg/l, mesmo que por apenas algumas horas, podem resultar em grandes perdas de peixes ([5]). Assim, o estudo da evolução temporal da concentração do OD em diversos locais de um rio, bem como a predição de valores futuros são contributos importantes na monitorização da qualidade da água e na prevenção da poluição da mesma. Este trabalho apresenta um estudo comparativo entre vários modelos competivos para descrever o comportamento mensal do OD na estação de monitorização de qualidade do Carvoeiro no rio Vouga. Uma descrição da bacia hidrográfica e caracterização das suas estações de monitorização pode ser encontrada no trabalho de [2].
O período usado na modelação compreende os meses de janeiro de 2002 a Maio de 2013 e os dados foram recolhidos do portal SNIRH ([6]). Estes modelos podem ainda ser avaliados quanto ao seu desempenho na predição de valores futuros do OD.

2. MODELOS EM ANÁLISE

Como a série da concentração de OD tem um comportamento periódico, os modelos que se apresentam são definidos tendo em conta essa característica da variável em estudo. O modelo base, que é comum aos modelos em estudo, é um modelo linear que tem em conta a variação sazonal na concentração de OD durante o ano, bem como a possibilidade da taxa de variação poder variar de acordo com o mês isto é, considera doze declives e doze ordenadas na origem, podendo ser escrito como

$$\begin{aligned} Y_t &= (\alpha_1 \cdot t + \beta_1) I_{t,1} + (\alpha_2 \cdot t + \beta_2) I_{t,2} + ... + (\alpha_{12} \cdot t + \beta_{12}) I_{t,12} + \xi_t \\ &= \sum_{s=1}^{12} (\alpha_s \cdot t + \beta_s) \cdot I_{t,s} + \xi_t, \end{aligned} \tag{1}$$

onde Y_t é a concentração do OD no mês t, α_s e β_s, $s = 1, 2, ..., 12$, são respetivamente o declive e a ordenada na origem associados ao mês $t = s + 12k$, para algum $k = 0, 1, 2, ...$ A função indicatriz $I_{t,s}$ é definida por $I_{t,s} = 1$ if $t = s + 12k$, para algum $k = 0, 1, 2, ...$ e $I_{t,s} = 0$ caso contrário, e ξ_t é um ruído branco ($E(\xi_t) = 0$, $var(\xi_t) = \sigma_\xi^2$ e $E(\xi_t \xi_r) = 0$ for $t \neq r$).
A estimação dos parâmetros do modelo, usando o método dos mínimos quadrados, revelou que, para o local em análise, os resíduos apresentavam uma correlação moderada. Adicionalmente, a análise dos resultados da estimação do modelo definido em (1), e após a correção dos erros padrão pelo facto dos erros serem correlacionados (ver [1]), apenas o declive associado ao mês de maio apresentou um valor estatisticamente significativo (para uma significância de 10%) pelo que o modelo base final possui 13 parâmetros, $\alpha_5, \beta_1, \cdots, \beta_{12}$. Pretende-se então usar modelos que consigam captar melhor a correlação existente na série temporal do OD que têm na sua estrutura o modelo base de regressão com 13 parâmetros e a inclusão de outras componentes estocásticas que permitirão descrever melhor a série em estudo.

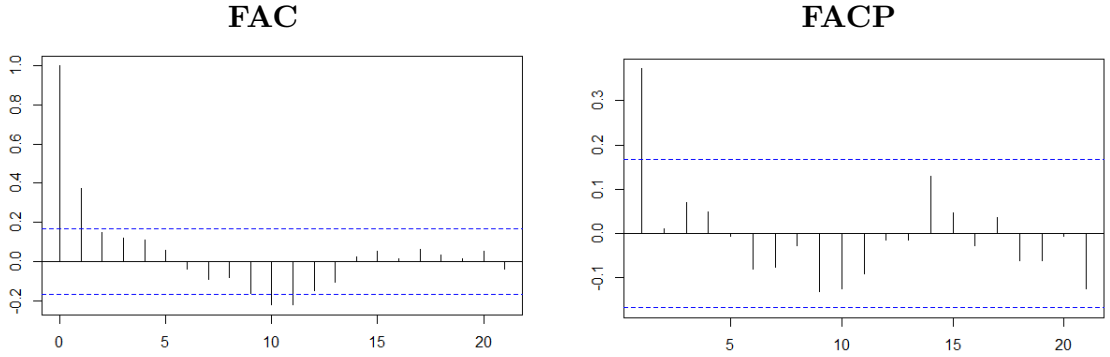


Figura 1: FAC e FACP empírica dos resíduos do modelo de regressão definido em (1).

2.1 Modelo de regressão linear com erros correlacionados (MI)

Tendo em conta que as funções empíricas de autocorrelação e de autocorrelação parcial que se apresentam na Figura 1, podemos definir o modelo linear com erros correlacionados, seguindo uma estrutura autorregressiva de ordem 1, descrito por

$$Y_t = \alpha_5 \cdot t \cdot I_{t,5} + \sum_{s=1}^{12} \beta_s \cdot I_{t,s} + \xi_t \tag{2}$$

$$\xi_t = \rho \xi_{t-1} + a_t, \tag{3}$$

com $\{a_t\}$ um ruído branco cuja variância é σ_a^2.

A estimação dos parâmetros deste modelo pode ser efetuada por decomposição, estimando-se inicialmente os 13 parâmetros associados ao modelo base (1), pelo método dos mínimos quadrados, e posteriormente são estimados, pelo mesmo método, os parâmetros da equação (3), usando os resíduos obtidos na primeira etapa de estimação.

2.2 Modelo em espaço de estados aditivo (MII)

Atendendo a que a variabilidade associada ao processo autorregressivo no modelo MI poderá não ser suficiente para explicar toda a variabilidade da série de OD, considera-se o modelo mais geral, que contém duas fontes de variabilidade. O modelo de espaço de estados é definido por

$$\begin{aligned} Y_t &= \alpha_5 \cdot t \cdot I_{t,5} + \sum_{s=1}^{12} \beta_s \cdot I_{t,s} + \xi_t + \epsilon_t & (4) \\ \xi_t &= \rho\xi_{t-1} + a_t. & (5) \end{aligned}$$

As equações (4)– (5) são respetivamente as equações de observação e de estado. Na equação de observação, os valores observados dependem de uma componente conhecida, de um fator estocático aditivo $\{\xi_t\}$ e uma componente de erro aleatória, $\{\epsilon_t\}$, designada por processo dos erros de observação que são descritos por um ruído branco com variância σ_ϵ^2. A equação do estado descreve o comportamento da componente estocástica, $\{\xi_t\}$, que se assume ser um processo autorregressivo de ordem 1 estacionário de média zero e parâmetro autorregressivo $|\rho| < 1$; o erro do estado, $\{a_t\}$ é não correlacionado com o processo dos erros de observação ($E(a_t\epsilon_r) = 0, \forall t, r$) e também descrito por um ruído branco. Adicionalmente os processos de ambos os erros são considerados ter distribuição normal. De notar que o modelo MI é um caso particular deste modelo quando se admite que a variância associada ao erro de observação é nula.

2.3 Modelo de calibração (MIII)

O terceiro modelo possui também duas componentes principais numa estrutura multiplicativa. A primeira é a componente base de regressão e a segunda componente estocástica que mensalmente irá calibrar a primeira componente. O modelo acima descrito é um modelo em espaço de estados, que pode ser definido por

$$\begin{aligned} Y_t &= X_t \cdot \left(\alpha_5 \cdot t \cdot I_{t,5} + \sum_{s=1}^{12} \beta_s \cdot I_{t,s} \right) + e_t & (6) \\ X_t &= \mu + \phi(X_{t-1} - \mu) + \varepsilon_t. & (7) \end{aligned}$$

Este modelo difere do anterior por ter uma estrutura multiplicativa, em que processo do estado $\{X_t\}$ tem uma estrutura autorregressiva de ordem 1 de média não nula μ. Também neste modelo os processos dos erros de observação e de estado são ruídos brancos e não correlacionados entre si. A distribuição associada aos erros é assumida ser a distribuição Gaussiana. Nos modelos MII e MIII, devido à sua estrutura, a estimação dos parâmetros destes modelos podem ser realizada em duas etapas. Na primeira etapa, que é comum aos três modelos, são estimados os parâmetros do modelo de regressão base, que, numa fase posterior, são considerados como conhecidos e serão usados para estimar os parâmetros associados ao estado. Nesta segunda etapa, como o estado é não observável e os modelos são lineares em que se assume a normalidade dos erros de observação e de estado, a estimação dos parâmetros é realizada através do método de máxima verosimilhança em que a predição a um passo dos valores observados de Y (ou de uma série obtida a partir desta) é obtida pelo algoritmo do filtro de Kalman ([4],[3]). Para a obtenção das estimativas dos parâmetros de máxima verosimilhança é necessário a utilização de procedimentos numéricos disponíveis em diversos *softwares*.

3. ALGUNS RESULTADOS

A Tabela 1 resume os resultados da modelação do OD na estação do Carvoeiro, rio Vouga no período em análise. Os erros quadráticos médios dos três modelos estão próximos, assim como os coeficientes de determinação. Os modelos MI e MII têm um desempenho ligeiramente superior ao modelo MIII. De salientar que a estimativa da variância do erro de observação do modelo MII é estatisticamente significativa, diferenciando assim os dois modelos MI e MII. No que concerne à análise dos resíduos dos modelos em estudo, estes não apresentam correlação significativa e os histogramas não estão visualmente longe da normalidade. No entanto apenas no modelo MII o teste de Kolmogorov-Smirnov não rejeita a normalidade dos resíduos que é um pressuposto comum a todos os modelos.

Tabela 1: Resultados da estimação dos parâmetros dos modelos

Modelo Base			MI			MII			MIII		
par.	est.	err. pad.*	par.	est.	err. pad.	par.	est.	err. pad.	par.	est.	err. pad.
α_5	-0.010	0.005									
β_1	10.209	0.25									
β_2	9.983	0.24									
β_3	9.546	0.24									
β_4	8.975	0.44									
β_5	9.262	0.24									
β_6	7.771	0.25									
β_7	8.164	0.25									
β_8	8.844	0.25									
β_9	7.941	0.25							μ	1.000	0.001
β_{10}	7.844	0.25	ρ	0.311	0.08	ρ	0.4120	0.02	ϕ	0.594	0.016
β_{11}	8.968	0.25	σ_a^2	0.594	–	σ_a^2	0.452	0.03	σ_ε^2	0.003	0.16×10^{-3}
β_{12}	9.946	0.25				σ_e^2	0.064	0.02	σ_e^2	0.266	0.011
R^2 0.561			R^2 0.598			R^2 0.602			R^2 0.595		
E.Q.M. 0.579			E.Q.M. 0.528			E.Q.M. 0.528			E.Q.M. 0.534		

AGRADECIMENTOS

Os autores foram parcialmente financiados por fundos portugueses através do CIDMA e da FCT–Fundação para a Ciência e a Tecnologia, dentro do projeto UID/MAT/04106/2013.

Referências

[1] Alpuim, T. e El-Shaarawi, A. (2008) On the efficiency of regression analysis with AR(p) errors. Journal of Applied Statistics, 35:7, 717–737.

[2] Costa, M; Monteiro, M. (2016). Discrimination of water quality monitoring sites in River Vouga using a mixed-effect state space model. Stochastic Environmental Research and Risk Assessment 30, 2: 607–619.

[3] Costa, M, Monteiro, M. (2016). Bias-correction of Kalman filter estimators associated to a linear state space model with estimated parameters. Journal of Statistical Planning and Inference, 176: 22 - 32.

[4] Harvey AC. (2006). *Forecasting structural time series models and the Kalman filter*. Cambridge, Cambridge University Press.

[5] Shifflett DS. Water and Sustainability. `http://www.unc.edu/~shashi/TablePages/dissolvedoxygen.html` (acedido a 16 de julho de 2014).

[6] Portal da Água. Instituto da Água. I.P. (INAG). `http://portaldaagua.inag.pt` (acedido a 20 Março de 2018).

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

MODELAÇÃO CONJUNTA DE DADOS LONGITUDINAIS E DE SOBREVIVÊNCIA EM DESISTÊNCIA DA PSICOTERAPIA

Ângela Ferreira[1], Inês Sousa[1], Eugénia Ribeiro[2], Miguel Gonçalves[2] e Paulo Machado[2]

[1] Escola de Ciências, Universidade do Minho, Guimarães
[2] Escola de Psicologia, Universidade do Minho, Braga

RESUMO

A desistência da psicoterapia é um fenómeno frequente e problemático. As taxas de prevalência oscilam entre os 20% e 60% [1], [2], sendo que, na maioria das vezes, os clientes desistentes exibem piores resultados terapêuticos e demonstram um baixo nível de satisfação com o tratamento [3]. Para além de comprometer a eficácia da psicoterapia, a desistência do cliente acarreta sérios custos sócio económicos para o individuo, instituições e comunidade em geral, dada a subutilização dos recursos e a manutenção ou exacerbação dos problemas de saúde mental [1].

A investigação sobre desistência apresenta diversos problemas metodológico-analíticos que dificultam a sistematização, comparação e replicação de resultados. Em primeiro lugar, não existe uma definição única e consensual do conceito de desistência, pelo que os investigadores recorrem a diferentes operacionalizações do conceito [2]. Para além disso, verifica-se a utilização de técnicas de análise subótimas (ex. análise de regressão e análise da variância), que desconsideram a natureza longitudinal dos dados. Acresce a observação de que, frequentemente, as covariáveis são avaliadas uma única vez e/ou assumidas como estáticas, quando na verdade podem e devem mudar ao longo do tempo (ex. nível de sintomatologia e qualidade da relação terapêutica) [4]. Estas análises ignoram também o facto de que os dados completos observados são condicionais aos indivíduos não terem desistido.

O presente estudo tem por objetivo caracterizar o tempo até à desistência da psicoterapia e testar o valor preditivo da qualidade da aliança terapêutica. Para tal, utilizar-se-á um conjunto de dados de natureza longitudinal, recolhidos ao longo do processo psicoterapêutico de 103 casos clínicos. A recolha decorreu numa clínica universitária, entre setembro de 2015 e maio de 2018. Neste contexto, são avaliadas e tratadas diversas

problemáticas como perturbações de ansiedade e de humor, problemas de ajustamento e relacionais e perturbações de personalidade.

Numa primeira fase, proceder-se-á a análises de sobrevivência e longitudinais separadas. Posteriormente, proceder-se-á à modelação conjunta dos dados, adotando a metodologia de efeitos aleatórios e considerando a distribuição condicional e não marginal dos dados. Os resultados serão discutidos considerando as suas implicações para a prática clínica.

Palavras e frases chave: Desistência da psicoterapia, Modelação Conjunta, Sobrevivência e Dados Longitudinais.

Referências

[1] J. K. Swift e R. P. Greenberg, «Premature discontinuation in adult psychotherapy: A meta-analysis.», *J. Consult. Clin. Psychol.*, vol. 80, n. 4, pp. 547–559, 2012.

[2] M. Wierzbicki e G. Pekarik, «A meta-analysis of psychotherapy dropout.», *Prof. Psychol. Res. Pract.*, vol. 24, n. 2, pp. 190–195, 1993.

[3] S. Knox, N. Adrians, E. Everson, S. Hess, C. Hill, e R. Crook-Lyon, «Clients' perspectives on therapy termination», *Psychother. Res.*, vol. 21, n. 2, pp. 154–167, Mar. 2011.

[4] A. F. Corning e E. V. Malofeeva, «The application of survival analysis to the study of psychotherapy termination», *J. Couns. Psychol.*, vol. 51, n. 3, pp. 354–367, 2004.

III Encontro Luso-Galaico de Biometría
Aveiro, 28 - 30 June 2018

NIRS AS A RAPID SCREENING METHOD TO PREDICT FIBER CONTENT IN SUGARCANE

Gonçalves, M.T.V.[1], Cardoso, W. J.[2], Roque, J. V.[2], Ferreira, R. A[3]., Peternelli, L.A.[3]

[1] Universidade Federal Viçosa, Department of Genetics and Plant Breeding,Viçosa,Brazil
[2] Universidade Federal Viçosa, Department of Chemistry, Viçosa, Brazil
[3] Universidade Federal Viçosa, Department of Statistics,Viçosa, Brazil

ABSTRACT

Sugarcane is an important crop in the tropical and subtropical zone. Due its rusticity and efficient photosynthetic ability to convert and store energy obtained from sunlight into chemical molecules, it widely distinguishes and outperforms other energy crops currently exploited aiming the production of bioenergy [9]. In the landscape of environmental concerns, bioenergy crops have recently gained much attention, especially when it comes at replacing fossil fuels as a renewable source of energy [5]. Studies applying chemometric techniques coupled with near-infrared spectroscopy (NIRS) data have been done with several applications, remarkably in agriculture research and crop production industry. These results have proved the potential of the strategy at profiling and evaluating large sample sets [2,5,6]. In addition, the method has several attractive features, including speed, non-destructive and environmental friendly aspects. The initial steps of the Sugarcane Genetic Breeding Program (PMGCA) of the Federal University of Viçosa (UFV) are hindered by the lack of experimental space and massive number of clones generated after crossings [8]. Consequently, high-throughput methods need to be developed as an attempt to adopt a more effective strategy and eventually circumvent these difficulties. Sugarcane clones with high rates of fiber on its constitutional tissues are the more likely candidates to be used in commercial plantations if the goal is to harness its biofuel production potential. In this work, we attempted to ascertain whether it is possible to develop a multivariate calibration model to predict fiber content in sugarcane clones using dry bagasse samples.

The spectral data matrix (Figure 1a) consisted of 141 samples of dry grind sugarcane bagasse, each representing a different genotype and was obtained from a Fourier Transform Near Infrared Spectrometer (FT-NIR) in an investigated range of 10 000–4000 cm^{-1}. Sample scans were the result of 32 averaged scans and were measured using the reflectance mode (log 1/R). The samples were collected during February 2017, in a collection at the experimental field of the UFV plant science department, Viçosa, Minas Gerais (MG), Brazil. Samples were derived from the bottom third of each stalk, taken to the laboratory and smashed with a hydraulic press. The remaining bagasse was dried for 24 hours at 50 ºC and grinded. Fiber content determination followed the method described by Legendre [4], and multivariate analysis was performed using MATLAB (2016a) software (Math Works, Natick, USA). The data was split into calibration and validation sets using the Kennard-Stone algorithm [3]. The NIR spectra matrix was pre-treated aiming the increase of signal to noise ratio, reduction of baseline variation, dimensionality and collinearity

and therefore improve the accuracy of the model. Different pre-treatments were tested and, based on the RMECV value, the following pre-treatments were chosen: smoothing with Savitzky-Golay algorithm [7], multiplicative scatter correction (MSC), mean centering and 1st derivative also with Savitzky-Golay algorithm (Figure 1b).

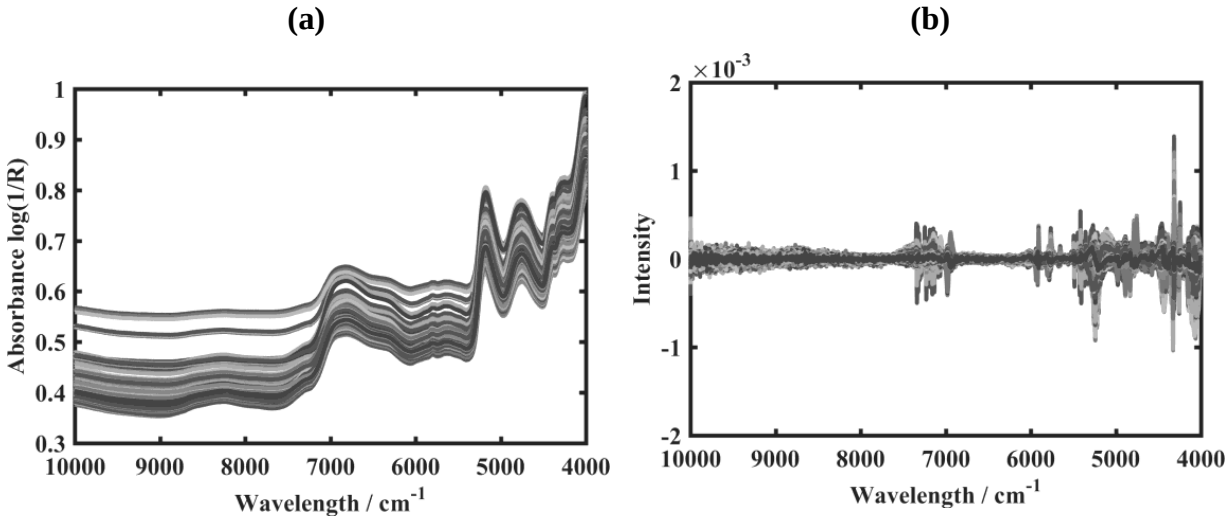


Figure 1. a) Raw NIR spectra and b) pre-treated NIR spectra.

The association between the NIR spectra and fiber content experimental values were done by building a multivariate calibration model using partial least squares regression (PLS) with eight latent variables. The model showed a correlation coefficient of cross-validation (Rcv) of 0.9021, and a root mean square error of cross-validation (RMSECV) of 3.3409 (Table. 1). The model ability to predict external sample sets was determined to predict the validation set and returned a prediction correlation coefficient (Rp) of 0.8856 and a root mean square error of prediction (RMSEP) of 4.1131 (Table.1).

Table 1. Statistical parameters of NIR-PLS model

RMSECV	3.3409
RMSEP	4.1131
R_{cv}	0.9021
R_p	0.8856

RMSECV: root mean squared error of cross-validation; RMSEP: root mean square error of prediction; R_{cv}: correlation coefficient of cross-validation; R_p: correlation coefficient of prediction.

A linear correlation between measured and predicted values can be seen in Figure 2.

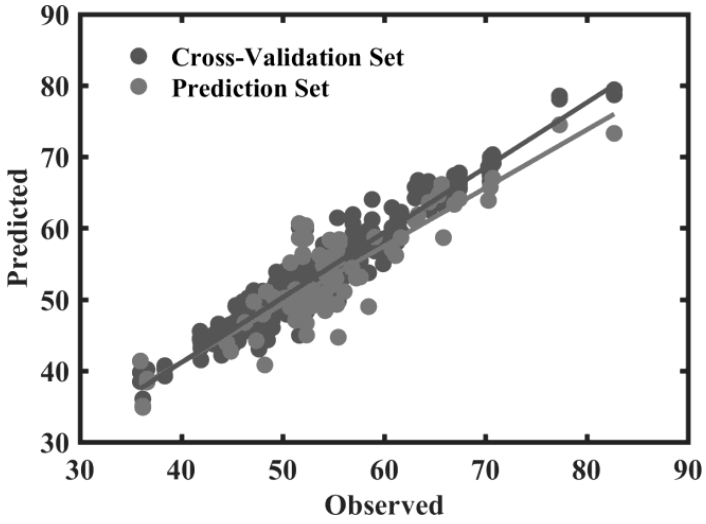


Figure 2. Observed *versus* predicted samples for the cross-validation set and prediction set.

Although the model presents high values of RMSECV and RMSEP, the respectively relative errors, as shown in Figure 3 (a) and (b), are mostly below 10%, suggesting the model could be used to predict unknown samples in future samplings. Other chemometric methods could be used to improve the model's prediction potential, as outlier detection and variable selection, not performed herein.

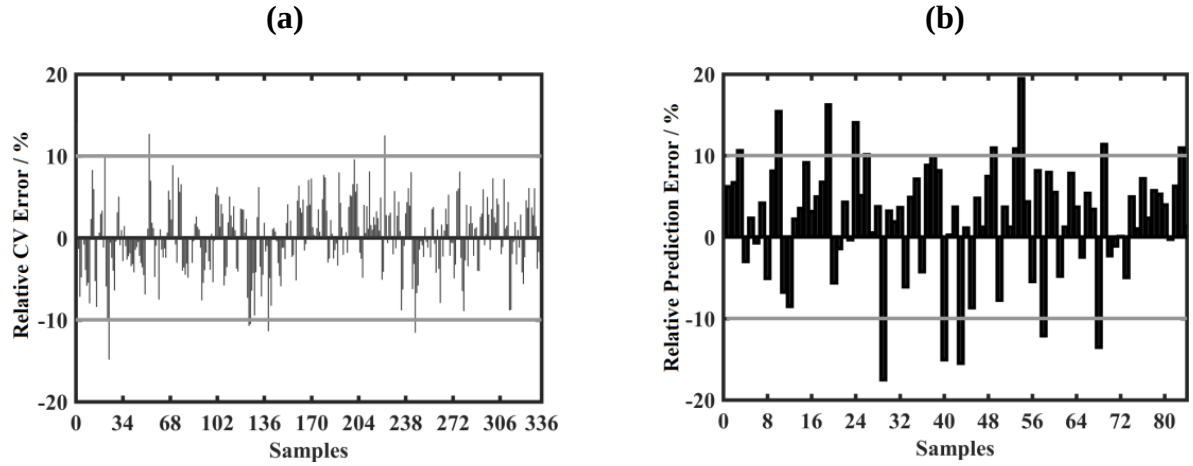


Figure 3. a) Relative cross-validation error of samples and b) relative prediction error of samples from the PLS model.

In this study, we examined the feasibility of building a calibration model to predict fiber content in sugarcane clones. Although future research still has to be conducted to assure whether NIRS can be used as a reliable phenotyping method, reasonably results were obtained, in which we may infer that dry, grind and therefore more homogeneous samples provide good results and can be used for prediction purposes.

Keywords and key sentences: NIR, PLS, Sugarcane, Fiber,

ACKNOWLEDGMENT

The authors would like to acknowledge the financial supporters of this research, the Inter-University Network for the Development of the Sugar and Ethanol Industry (Ridesa) and the National Council of Technological and Scientific Development (CNPQ).

References

[1] Assis, C. & Ramos, R. S. & Silva, L. A. & Kist, V. & Barbosa, M. H. P. & Teófilo, R.F. (2017) Prediction of lignin content in different parts of sugarcane using near-infrared spectroscopy(NIR), ordered predictors selection(OPS) and partial least squares(PLS). *Applied Spectroscopy,* 0(0), 1-12.

[2] Belini, U. L. & Hein, P. R. G & Filho M. T. & Rodrigues J. C. & Chaix, G. (2011) Near-infrared spectroscopy for estimating sugarcane bagasse content in medium density fiberboard. *BioResources.*6(2), 1816-1829.

[3] Kennard, R. W., & Stone, L.A. (1996). Computer-aided design of experiments. *Technometrics*, 11(1), 137-148

[4] Legendre, B. L. & Burner, D. M. (1995) Biomass Production of Sugarcane Cultivars and Early-generation Hybrids. *Biomass and Bioenergy* 8, 55-61.

[5] Monrroy, M. & Garcia, J. R. & Troncoso, E. & Freer, J. (2014) Fourier transformed near-infrared(FT-NIR) spectroscopy for the estimation of parameters in pretreated lignocellulosic materials for bioethanol production. *Journal of Chemical Technology and Biotechnology*.90(7), 1281-1289.

[6] Roque, J. V. & Dias, L. A. S. & Teófilo, R. F. (2017). Multivariate calibration to determine phorbol esters in seeds of *Jatropha curcas L*. using near infrared and ultraviolet spectroscopies. *J.of Braz. Chem. Soc*. 28, 1506-1516.

[7] Savitzky. A & Golay, M. J. E. (1964) Smoothing and differentiation by simplified least squares procedures, *Anal. Chem.*36,1627-1639.

[8] Silveira, L. C. I. & Brasileiro, B. P. & Weber, H. & Daros, E. & Peternelli, L.A. & Barbosa M.H.P. (2016). Selection in energy cane families. *Crop Breeding and Applied Biotechnology*, 16,298-306.

[9] Tew, T. L. & Cobill R. M. (2008) Genetic improvement of sugarcane (Saccharum spp.) as an energy crop. *Genetic Improvement of Bioenergy crops*.208, 273-294.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

Análise de risco na atividade florestal

Mónica Rodrigues[1], Maria da Conceição Costa[1,2] e Isabel Pereira[1,2]

[1]Departamento de Matemática, Universidade de Aveiro, Aveiro, Portugal
[2]CIDMA, Universidade de Aveiro, Aveiro, Portugal

RESUMO

Atualmente, a atividade florestal e a cadeia de produtividade a ela aliada assumem um importante papel na economia de Portugal, tornando-se crucial a formulação de estratégias e instrumentos que a apoiem.

Considera-se neste trabalho o indicador de produtividade florestal Acréscimo Médio Anual em Volume (AMAV, m^3/ha/ano) que suporta diversos processos de decisão em planeamento e gestão florestal (idade de corte, seleção do modelo de silvicultura, exploração florestal).

A modelação da produtividade florestal baseia-se em medições que refletem as condições médias que ocorreram no período de tempo em que as medições sucederam. As alterações climáticas e outros eventos (direta ou indiretamente relacionados com estas alterações) como o aumento da ocorrência de pragas e doenças ou do risco de incêndio traduzem-se em maior incerteza na obtenção de estimativas de produtividade e na tomada de decisão florestal, [1],[3].

Pretende-se neste estudo analisar de que forma o risco e a incerteza na ocorrência de uma das pragas que mais danos causa em povoamentos de eucalipto, o gorgulho do eucalipto, poderá afetar a estimativa da produtividade florestal em regiões de risco fraco a muito forte.

No desenvolvimento da presente investigação, objetivando-se a implementação de um modelo que permita dar resposta ao problema colocado, foi feita uma análise de regressão linear múltipla, com inclusão de variáveis *dummy*, [2]. A análise do modelo construído permitiu detetar nos resíduos a presença de heterocedasticidade e autocorrelação. Face à problemática referida, foi necessário aplicar métodos estatísticos adequados, nomeadamente métodos de regressão robusta tais como regressão linear múltipla robusta (RLMR) e métodos baseados em estimadores consistentes na presença de heterocedasticidade e autocorrelação (HAC - *heteroskedasticity and autocorrelation consistent*), [4].

Palavras-chave: Análise de risco, autocorrelação, heterocedasticidade, produtividade florestal, regressão robusta, variáveis *dummy*.

Referências

[1] Branco, M., Grodzi, W., Jacquet, J. S., Moreira, F., Netherer, S., Schelhaas, M. J., Tomé, M. (2011). Report on specific risk analysis in regional forests of Europe under various Forest Management Alternatives. *Report on specific risk analysis in regional forests of Europe under various Forest Management Alternatives, EFI Technical Report 67, European Forest Institute*, 19-25.

[2] Mendes de Oliveira, M., Santos, L. D., Fortuna, N. (2011). *Econometria*. Escolar Editora, Lisboa.

[3] Reis, A. R., Ferreira, L., Tomé, M., Araujo, C.,Branco, M. (2012). Efficiency of biological control of Gonipterus platensis (Coleoptera: Curculionidae) by Anaphes nitens (Hymenoptera: Mymaridae) in cold areas of the Iberian Peninsula: implications for defoliation and wood production in Eucalyptus globulus. *Forest Ecology and Management*, 270, 216–222.

[4] Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11(10), 1–17.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

ANÁLISE DA PRESENÇA DE VARIÁVEIS MEDIADORAS - APLICAÇÃO A DADOS DE UM INQUÉRITO REALIZADO NA CIDADE DA PRAIA EM CABO VERDE

Catarina Venda[1], P. de Zea Bermudez[2] e Luzia Gonçalves[3]

[1] Faculdade de Ciências da Universidade de Lisboa
[2] CEAUL e Faculdade de Ciências da Universidade de Lisboa
[3] CEAUL e GHTM/Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa

RESUMO

A mediação desempenha um papel muito importante em certas áreas como a Psicologia e as Ciências da Saúde [3,4,5].

A mediação indica se a relação entre uma variável dependente Y e uma variável independente X pode ser explicada pela intervenção de outra variável intermédia (mediação simples) ou por várias variáveis intermédias (mediação múltipla), denominadas variáveis mediadoras. Este conceito é frequentemente referido como mediação causal, pois tem implícito o conceito de causalidade: a variável independente causa a variável mediadora, a qual, por sua vez, causa a variável dependente. O presente estudo apresenta o "estado da arte" relativamente às principais metodologias utilizadas na avaliação da mediação simples. Entre as metodologias principais são referidas as abordagens tradicionais e a abordagem contrafactual. A abordagem contrafactual baseia-se na teoria dos resultados potenciais desenvolvida no contexto da inferência causal [2].

A aplicação destas metodologias é ilustrada com dados recolhidos no âmbito do projecto de investigação UPHI – STAT (PTDC/ATP-EUR/5074/2012) realizado na cidade da Praia, Cabo Verde. Os dados foram recolhidos através da aplicação de um inquérito a habitantes de três zonas da cidade da Praia que apresentam diferentes características, com o objectivo de analisar as suas desigualdades em termos de Saúde [1]. Escolhendo algumas variáveis do questionário, são analisadas três situações que têm em comum as variáveis independente e dependente, variando em cada situação analisada, a potencial variável mediadora. Todas as variáveis foram tratadas como binárias. O objectivo é concluir quanto à presença ou ausência de mediação, através da estimação dos efeitos de interesse neste contexto – efeitos directo, indirecto e total. A estimação dos efeitos é realizada com recurso ao pacote "*mediation*" do *R*, ao qual está subjacente a abordagem contrafactual [6].

Palavras e frases chave: Mediação, causalidade, efeitos directos e indirectos, abordagem contrafactual.

AGRADECIMENTOS

Este trabalho foi financiado pela FCT - Fundação para a Ciência e a Tecnologia, Portugal, através dos projectos UID/MAT/00006/2013 e PTDC/ATP-EUR/5074/2012.

Referências

[1] Gonçalves, L., et al. (2015). Urban Planning and Health Inequities: Looking in a Small-Scale in a City of Cape Verde, *PLoS ONE*, 10(11): pp 1-27.

[2] Imbens, G. W., Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.

[3] MacKinnon, D. P. (2008). Introduction to Statistical Mediation Analysis, Taylor & Francis Group, New York.

[4] MacKinnon, D. P., Fairchild, A. J., Fritz, M. S. (2007). Mediation Analysis, *Annual Review of Psychology*, 58: pp 593–614.

[5] Preacher, K. J., Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models, *Behavior Research Methods, Instruments, & Computers*, 36 (4): pp 717-731.

[6] Tingley, D., Yamamoto, T., Hirose, K., Keele, L., Imai, K. (2014). mediation: R Package for Causal Mediation Analysis: pp 1-40.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

COMPARAÇÃO BAYESIANA DE TESTES DE DIAGNÓSTICO COM DADOS DENSAMENTE OMISSOS AO ACASO

Carlos Daniel Paulino[1] e Giovani L. Silva[2]

[1]Centro de Estatística e Aplicações (CEAUL) & IST, Universidade de Lisboa
[2]Dep. Matemática, Instituto Superior Técnico (IST) & CEAUL, Universidade de Lisboa

RESUMO

Este trabalho é uma sequela de um artigo (Poleto *et al.*, 2011) sobre comparação de testes, assente num conhecido padrão de ouro, através das usuais medidas de acurácia por meio de métodos frequencistas, num quadro de substancial omissão de dados segundo um processo não informativo. Esta sequela passa a adotar uma abordagem bayesiana por se entender mais adequada para lidar com a incompletude do grosso dos dados, sem recurso a argumentos apoiados em grandes amostras. Computacionalmente, esta análise recorre a eficientes métodos de Monte Carlo iterativo com restringida ampliação de dados, executados num programa *ad hoc* em R. Em cada passo *a posteriori*, após fácil simulação de apropriadas variáveis latentes, o parâmetro de interesse é simulado diretamente à custa de distribuições Dirichlet, pela caraterização da generalização destas em padrões monótonos de dados.

Palavras-chave: Omissão ao acaso, medidas de acurácia do diagnóstico, Inferência bayesiana, Distribuição Dirichlet generalizada, Algoritmo de ampliação de dados em cadeia.

Referências

[1] Dickey, J.M., Jiang, J.M., Kadane, J.B. (1987). Bayesian methods for censored categorical data. *Journal of the American Statistical Association* 82, 773–781.

[2] Paulino, C.D., Amaral Turkman, A., Murteira, B., Silva, G. (2018). *Estatística Bayesiana*, 2ª edição. A sair em breve.

[3] Poleto, F.Z., Singer, J.M., Paulino, C.D. (2011). Comparing diagnostic tests with missing data. *Journal of Applied Statistics* 38, 1207–1222.

[4] Tanner, M.A., Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* 82, 528–550.

[5] Turkman, M.A., Paulino, C.D. (2015). *Estatística Bayesiana Computacional – uma introdução*. Edições SPE, Lisboa.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

UMA BASE CONCEITUAL RACIONAL PARA O EXPERIMENTO

João Gilberto Corrêa da Silva [1]

[1] Universidade Federal de Pelotas, Pelotas, RS, Brasil (Professor Titular – aposentado).

RESUMO

Textos e ensino de Estatística Experimental enfatizam a análise estatística de experimentos e consideram superficialmente sua fundamentação conceitual. Conceitos básicos são definidos de modo impreciso, incoerente e ambíguo. Esse é o caso, por exemplo, dos conceitos de material experimental, fator experimental, unidade experimental e erro experimental. As consequências são incompreensão e emprego incorreto desses conceitos, falhas no planejamento e análise de experimentos e, por decorrência, ineficiência de muitas pesquisas e desperdiço de recursos. Esta comunicação faz uma revisão e reformulação de conceitos importantes com o propósito de estabelecer uma base conceitual racional para o experimento.

Palavras e frases chave: Estatística Experimental, material experimental, fator experimental, unidade experimental, fator de unidade, erro experimental.

1. INTRODUÇÃO

A precariedade dos conceitos formulados na literatura é notável mesmo nos textos consagrados. As seguintes definições de unidade experimental são ilustrativas: “quantidade total de material a qual um tratamento é aplicado em uma repetição simples” (Federer, 1955, p. 58); “grupo de material ao qual é aplicado um tratamento em um ensaio simples do experimento” (Cochran & Cox, 1957, p.15); “A definição formal de uma unidade experimental é que ela corresponde a menor divisão do material experimental tal que quaisquer duas unidades podem receber tratamentos diferentes no presente experimento” (Cox, 1958, p.2); “parte de material experimental a que um tratamento é assinalado e aplicado” (Hinkelmann & Kempthorne, 1994, p.36). Essas definições são imprecisas, em parte por dependerem de conceitos, como material, material experimental, tratamento e repetição que não são definidos.

A dificuldade das definições de conceitos básicos é admitida por Bailey (2008, p. 8) quando enuncia as seguintes definições de unidade experimental e de tratamento, que ela reconhece serem circulares: "uma unidade experimental é a menor unidade à qual pode ser

aplicado um tratamento", "um tratamento é a descrição completa do que pode ser aplicado a uma unidade experimental".

A insatisfação com essas definições é manifestada na literatura. Brien & Demétrio (1998) salientam a diversidade de opiniões que têm sido expressas sobre a identificação da unidade experimental em experimentos de pastoreio com gado de corte e comentam que é controverso se a unidade experimental é o animal individual, a unidade de campo ou a unidade de campo e seus animais. Eles afirmam que o reconhecimento de que o erro que afeta efeitos de tratamentos envolve a variabilidade de campo e de animal e o desejo de que este erro seja determinado pelas unidades experimentais motivam a declaração de que a unidade experimental é a unidade de campo e os animais que ela suporta.

Entretanto, essa não é a única origem desse erro. Ele também resulta dos efeitos das características do ambiente, relacionadas com clima, doenças, pragas, invasoras e predadores, das técnicas de manejo e dos processos de mensuração. A caracterização completa da unidade experimental requer a descrição de todas as características da amostra que lhe correspondem. Essa abrangência do conceito de unidade experimental é reconhecida por Cox, quando complementa sua definição: "Consideramos serem incluídos na definição da unidade todos os aspectos da configuração do experimento não envolvidos no tratamento, isto é, aqueles que são independentes da assinalação particular dos tratamentos que é adotada." (Cox, 1958, p.191).

Silva (1999, 2004, 2008) propõe uma base conceitual para o experimento, racional e coerente com os significados reais. O objetivo da presente comunicação é apresentar uma síntese dessas contribuições.

2. BASE CONCEITUAL

O experimento é um método de pesquisa científica para inferências sobre relações causais de características das unidades de uma população objetivo, ou seja, de um subconjunto de características que exprimem o desempenho dessas unidades (*características respostas*) com um subconjunto de características que supostamente as afetam (*características explanatórias*) na presença das demais características dessas unidades (*características estranhas*). A definição dessas três classes de características é estabelecida pelo problema e a hipótese científica, que determinam os objetivos do experimento.

As inferências são baseadas na verificação da relação causal postulada em uma amostra da população objetivo. As unidades da amostra compreendem as mesmas três classes de características da população objetivo. Assim como as unidades da população objetivo, as unidades da amostra são sistemas complexos de características que interagem dinamicamente no espaço e no tempo.

Os conceitos básicos da pesquisa experimental devem considerar essa complexidade. Essa abordagem requer a identificação das três classes de características da amostra e uma descrição abrangente das características estranhas que assegure a identificação das características cujos efeitos possam ser confundidos de modo relevante com efeitos das características explanatórias. Essas três classes de características presentes na amostra constituem o *material experimental*. A fração do material experimental onde é efetuada uma observação independente de uma característica resposta é a *unidade de observação* desta característica. Os valores de uma característica resposta mensurados nas unidades de observação têm duas origens: características explanatórias e características estranhas. Os efeitos das características estranhas sobre as características respostas constituem o *erro experimental*. Os efeitos das características

explanatórias sobre características respostas são confundidos com o erro experimental. O recurso para reduzir esse confundimento e torná-lo não tendencioso é o *controle experimental*, que é efetuado por controle de técnicas experimentais, controle local, controle estatístico e casualização.

O planejamento do experimento define essas três classes de características e as relações entre as características explanatórias, entre as características estranhas e entre estas duas classes de características que constituem, respectivamente, a *estrutura das condições*, a *estrutura das unidades* e a *estrutura do experimento*. As especificações dessas estruturas estabelecem o *delineamento do experimento*. A estrutura das condições deve ser estabelecida em consonância com os objetivos do experimento, enquanto a estrutura das unidades é elaborada segundo a disponibilidade de material experimental.

O planejamento da estrutura das condições compreende as escolhas das características explanatórias ou *fatores experimentais*, dos níveis desses fatores e das combinações desses níveis, que são denominadas *condições experimentais* ou, simplesmente, *condições*. A associação dos níveis de um fator experimental às unidades da amostra pode ser controlada pelo pesquisador, por atribuição aleatória, ou ser inerente a estas unidades. No primeiro caso, o fator é denominado *fator de tratamento*; no segundo, *fator intrínseco*. Níveis de fator de tratamento recebem a designação particular de *tratamentos*.

O planejamento da estrutura das unidades compreende a escolha da unidade de observação e as definições do controle local e da associação entre os níveis dos fatores experimentais e as unidades de observação. Esse planejamento determina classificações das unidades de observação. Cada uma dessas classificações corresponde a uma característica estranha relevante, que constitui um *fator de unidade*. As classes de cada uma dessas classificações são os *níveis* do correspondente fator de unidade. O efeito de um fator de unidade é um componente do erro experimental.

A estrutura das condições e a estrutura das unidades são associadas pela casualização dos níveis dos fatores de tratamento às unidades de observação e a manifestação dos níveis dos fatores intrínsecos nessas unidades. A relação dessas duas estruturas constitui a *estrutura do experimento*. Assim, na estrutura do experimento há uma associação entre os fatores experimentais e os fatores de unidade e uma correspondência entre os níveis desses fatores. Os níveis de um fator de unidade são as *unidades experimentais* do fator experimental com o qual ele é associado. O número de unidades experimentais com uma nível de um fator experimental é o *número de repetições* deste nível.

A relação entre um fator experimental e um fator de unidade pode compreender mais de uma unidade experimental para seus níveis ou uma única unidade experimental para cada um de seus níveis. Na segunda situação os fatores são designados *equivalentes* ou *parceiros*. Ela ocorre quando o fator experimental é fator intrínseco ou fator de tratamento com uma repetição para cada um de seus níveis. Efeitos de fatores equivalentes são completamente confundidos. Essa propriedade é altamente relevante e deve ser considerada no planejamento do experimento.

Os fatores de unidade estratificam o conjunto das características estranhas do material experimental. Como consequência, o erro experimental é decomposto em tantos estratos quantos são os fatores de unidade. A fração do erro experimental que corresponde a um fator de unidade constitui um *estrato do erro experimental*. O *erro experimental que afeta um efeito de fator experimental* é uma fração do erro experimental composta por um subconjunto de seus estratos.

É conveniente que os planejamentos da estrutura das condições e da estrutura das unidades sejam procedidos separadamente. Esse procedimento é recomendável para que a estrutura do experimento seja expressa corretamente, particularmente em experimentos complexos.

Entretanto, a estrutura das condições é condicionada à disponibilidade de material experimental e a estrutura das unidades deve ser apropriada para a estrutura das condições. Nessas circunstâncias, essas duas estruturas são altamente interdependentes. Uma estratégia racional para a geração do delineamento experimental compreende a seguinte sequência de passos: 1) elaborar a estrutura das condições tendo em conta as restrições de material experimental; 2) considerar as estruturas de unidades alternativas para essa estrutura de condições; 3) escolher, entre essas estruturas de unidades, aquela que, associada à estrutura das condições, permita inferências mais eficientes sobre os efeitos dos fatores experimentais relevantes para os objetivos do experimento; 4) caso não seja encontrada uma estrutura de unidades satisfatória, reconsiderar a sequência de passos 1, 2 e 3. Os passos 1 e 2 podem conduzir à formulação de diversas estruturas de experimento. Como regra geral, o pesquisador deve escolher o delineamento experimental que proveja o máximo de informação relevante aos objetivos do experimento com o custo mínimo. Para tal devem ser levados em conta os princípios básicos do delineamento do experimento: repetição, controle local, casualização, ortogonalidade, balanceamento, confundimento e eficiência.

3. CONCLUSÕES

As definições dos conceitos básicos do experimento, precisas e coerentes com seus significados, e a especificação da estrutura do experimento pela derivação separada da estrutura das condições e da estrutura das unidades permitem a geração do delineamento experimental eficiente tanto sob o ponto de vista teórico como prático.

Referências

[1] Bailey R.A. (2008). *Design of comparative experiments*. Cambridge University Press, Cambridge, UK.

[2] Brien, C.J.; Demétrio, C.G.B. (1998) Using the randomization in specifying the ANOVA model and table for properly and improperly replicated grazing trials. *Australian Journal of Experimental Agriculture,* v.38, n.4, p.325-334.

[3] Cochran, W.G.; Cox, G.M. (1957). *Experimental design*. 2. ed. New York: John Wiley.

[4] Cox, D.R. (1958) *Planning of experiments*. New York: John Wiley.

[5] Federer, W. T. (1955) *Experimental design: theory and application*. New York: Macmillan,

[6] Hinkelmann, K.; Kempthorne, O. (1994) *Design and analysis of experiments*. v.1. New York: John Wiley.

[7] Silva, J.G.C. (2008). A conceptual basis and a new approach to the planning of experiments. In *Symposium on the Planning of designed experiments: Recent advances in methods and applications (DEMA2008)*. Isaac Newton Institute for Mathematical Sciences, University of Cambridge.

[8] Silva, J. G. C. (2004). A estrutura do experimento e o modelo estatístico. In *Estatística Jubilar - Actas do XII Congresso Anual da Sociedade Portuguesa de Estatística*, 735-744.

[9] Silva, J.G.C. (1999). A consideração da estrutura das unidades em inferências derivadas do experimento. *Pesquisa Agropecuária Brasileira* 34, 911-925.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

TEOR FOLIAR DE NUTRIENTES EM AMENDOIM (*Arachis hypogaea L.*) ASSOCIADOS COM FUNGOS MICORRÍZICOS ARBUSCULARES E SUPLEMENTADOS COM EXTRATO SOLÚVEL DE ALGAS, AVALIADOS POR ANÁLISE MULTIDIMENSIONAL "GLM E CANDISC"

Renata B. S. Coscolin[1], João R. Favan[1], Deoclécio Jardim Amorim [1], Edilson R. Gomes[2], Fernando Broetto[3] e Maria M. P. Sartori[1].

[1] Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agronômicas. Botucatu, Brasil
[2] Faculdades Integradas de Bauru (FIB). Bauru, Brasil
[3] Universidade Estadual Paulista (UNESP), Instituto de Biociências. Botucatu, Brasil

RESUMO

Do ponto de vista nutricional, as micorrizas arbusculares estão envolvidas em mecanismos que aumentam a absorção de nutrientes, como o fósforo. O objetivo do estudo foi avaliar os efeitos da aplicação de biofertilizante a base de extrato de algas (ESA) e da associação com fungos micorrízicos arbusculares (FMAs) em plantas de amendoim para o teor foliar de macro e micro nutrientes através da análise multidimensional. O experimento foi instalado em casa de vegetação usando o cultivar IAC Runner 886, com inoculação e sem inoculação (+ FMAs e –FMAs) com as seguintes espécies de FMAs (*Glomus intraradices* e *Glomus etunicatum*), além da suplementação ou não com o extrato solúvel de algas (+ ESA e –ESA). Foram avaliados nas folhas os teores de N, P, K, Ca, S, Mn, Fe, Z e B pela análise multidemental. Para a absorção de nutrientes a análise multidimensional demonstrou uma resposta sinérgica entre os FMAs e o ESA onde a variação na absorção das plantas de macro e micronutrientes demonstra que a suplementação com bioestimulantes é uma fonte nutricional viável que mantém a comunidade fúngica no solo ativa.

Palavras e frases chave: Macro e micro nutrientes, *Ascophyllun Nodosun*, variáveis canônicas.

1. INTRODUÇÃO

A cultura do amendoim (*Arachis hypogae L.)* tem despertado grande interesse econômico pois sua exploração comercial é uma opção favorável para agricultura devido as sementes constituirem uma importante fonte de proteína vegetal e de óleo comestível [1]. Existe uma diversidade de mercados para a cultura, seja para o consumo *in natura*, para a indústria oleoquímica ou produção de biodiesel [2].

O uso de bioestimulantes naturais, como o extrato de algas marinhas (ESA) provenientes da espécie *Ascophylun nodosun*, assim como a associação simbiótica com fungos micorrízicos (FMAs) está cada vez mais se difundindo na agricultura [3].

A adoção do manejo agronômico utilizando esses bioagentes podem estimular o crescimento e o desenvolvimento vegetal assim como a redução no fornecimento de fertilizantes [4]. Certos

elementos de pouca mobilidade do solo, como o fósforo, podem ser absorvidos mais eficientemetne pela planta [5] devido a expansão radicular promovida pela associação simbiótica entre as raízes das plantas e os fungos. O objetivo deste trabalho foi avaliar os efeitos da aplicação de biofertilizante e da associação com fungos micorrízicos em plantas de amendoim no teor foliar de macro e micro nutrientes pela análise multidimensional.

2. MATERIAL E MÉTODOS

A pesquisa foi desenvolvida em estufa agrícola localizada no instituto de biociências - UNESP, Campus de Botucatu-SP. Utilizou-se sementes do cultivar IAC- Runner 886 as quais foram cultivadas em vasos com capacidade de 30L em solo nas seguintes condições: T1: solo com inoculação de micorriza e suplementação com ESA (+FMAs + ESA), T2: solo com inoculação de micorriza e ausência da suplementação com ESA (+FMAs - ESA), T3: solo sem inoculação de micorriza e suplementação com ESA (-FMAs + ESA) e T4: solo sem inoculação de micorriza e ausência de suplementação com ESA (-FMAs – ESA).

A adição do ESA deu-se através do produto comercial Acadian ® conforme recomendação do fabricante sendo as iniciadas as aplicações aos 30 dias após o plantio (DAE) e as subsequentes realizadas semanalmente até o início da senescência das plantas (90 DAE). O procedimento de inoculação dos fungos foi em solo não estéril e utilizou-se uma mescla de 2 isolados, IAC 5 (*Glomus intraradicis*) e IAC 44 (*Glomus etunicatum*). Aplicou-se 100g do inócuo por vaso, junto a cova, no mesmo instante da semeadura.

A amostragem das folhas para análise nutricional priviligiou folhas normais, sadias e sem deficiência. O material vegetal foi levado para estufa de circulação forçada de ar a 60ºC, até peso constante sendo a seguir submetido a análises químicas.

Os teores de K foram determinados por fotometria de emissão de chamas; os de Ca,Mg,Zn, Cu, Fe e Mn por espcectrofotometria de absorção atômica; teores de P e S por espectrofotometria; e o teor de N pelo método Kjeldal segundo metodologia [6] .

O delineamento experimental foi inteiramente ao acaso e os dados referentes aos teores de macro e micronutrientes avaliados pela análise multidimensional por meio das técnicas de "GLM e CANDISC" utilizando o software SAS 9.0 [7].

3. RESULTADOS E DISCUSSÃO

Após a combinação das variáveis originais através da ACP identificou-se aquelas de maior variação sendo o N, P, Ca, Mg, B e Zn. Dada as variáveis de maior importância para o modelo as mesmas foram submetidas a uma análise de variância multidimensional, sendo esta significativa a 1% pelo teste de Wilks. A tabela 1 mostra as raízes (eigenvalues, em inglês) da matriz $E^{-1} H$ e, a partir delas obtivemos os coeficientes das variáveis canônicas.

Raízes	Valores das raízes *(eigenvalues)*	Proporção
λ_1	10,7907	0,7297
λ_2	3,4749	0,235
λ_3	0,5226	0,0353
$\lambda_1+\lambda_2+\lambda_3$	14,7882	100,00%

Tabela 1: Raízes da matriz $E^{-1} H$.

Observamos que cada raiz corresponde a uma variável canônica. No caso presente, a maior raiz (λ_1=10,7907) corresponde a aproximadamente 73% da variação, logo ela é capaz responder aos

objetivos dessa pesquisa. Ela nos dá a primeira variável canônica conforme visualizamos na Tabela 2.

Variáveis	CAN_1	CAN_2	CAN_3
N	-0,0152	0,0559	0,1595
P	-0,3422	2,2278	0,9887
Ca	0,0305	-0,4264	-0,3446
Mg	-0,1269	0,1316	0,9402
B	318,3864	-451,5960	55,3805
Zn	68,1855	16,6737	-10,1728

Tabela 2: Raízes da matriz $E^{-1}H$.

A primeira variável canônica aplicada aos valores de N, P, Ca, Mg, B e Zn de cada parcela do experimento, nos gerou uma variável única chamada aqui de Z. Foi aplicado a nova variável Z a análise de variância tradicional, unidimensional para qual o teste F de tratamentos tem o valor máximo. A análise de variância foi significativa, as médias da variável Z foram comparadas pelo teste de Scheffé (Tabela 3).

TRATAMENTO	MÉDIA DA VARIÁVEL Z
+FMA - ESA	21,9209 a
+FMA + ESA	19,9594 a
- FMA - ESA	15,6235 b
-FMA +ESA	15,2338 b

Médias comparadas pelo teste de Scheffé a 1% de probabilidade.

Tabela 3: Raízes da matriz $E^{-1}H$.

O procedimento CANDISC foi utilizado para realizar a análise canônica e derivar as respectivas funções canônicas pelas combinações lineares das variáveis quantitativas que resumem a variação entre os teores de macro e micro nutrientes avaliados em cada tratamento.

As variáveis canônicas obtidas apontaram efeitos positivos quando as plantas estão associadas com os FMAs e suplementadas ou não com ESA para a absorção de nutrientes.

A presença dos FMAs levaram as plantas à maior acúmulo de nutrientes quando comparadas as plantas que não tiveram a presença dos mesmo. A análise da compatibilidade hospedeiro e o comportamento da espécie vegetal, quando aplicado um bioestimulante como suplemento, indicam a ocorrência de sinergia entre as relações planta (hospedeiro), fungo e o bioestimulante para absorção de nutrientes.

O amendoim responde eficientemente a adubação [8] e a capacidade de resposta a colonização micorrízica está diretamente relacionada às características morfológicas e fisiológicas do hospedeiro. O material de amendoim estudado apresenta dependência micorrízica, ou seja, o crescimento e/ou produção dependente da colonização pelas FMAs em um determinado nível de fertilidade do solo [9]. Essa dependência micorrízica também pode estar associada a rusticidade, ou a baixa exigência nutricional da planta [10].

4. CONCLUSÕES

As plantas de amendoim associadas com fungos micorrizicos seja suplementadas ou não com ESA apresentaram maior eficiência na absorção de nutrientes. Significantes variações na

absorção de macro e micro nutrientes foram identificadas pela técnica “GLM e CANDISC” em plantas de amendoim.

Portanto, manejos, como a suplementação com bioestimulantes, que mantenham ou incrementem a comunidade nativa de fungos micorrízicos, pelo aporte nutricional mantendo assim a comunidade microbiana ativa do solo, seriam boas alternativas biotecnológicas para algumas espécies de leguminosas.

Referências

[1] Barbosa, R M.; Homem, B. F. M.; Tarsitano, M. A. A. (2014). Custo de produção e lucratividade da cultura do amendoim no município de Jaboticabal, São Paulo. *Revista Ceres* 61, 475–481.

[2] Correia, K. G. et al. (2009). Crescimento, produção e características de fluorescência da clorofila a em amendoim sob condições de salinidade. *Revista Ciência Agronômica* 40, 514–521.

[3] Augé, R. M. 2001. Water relations, drought and vesicular-arbuscular mycorrhizal symbiosis. *Mycorrhiza* 11, 3– 42.

[4] Van Oosten, M. J. et al. A. (2017). The role of biostimulants and bioeffectors as alleviators of abiotic stress in crop plants. *Technol. Agric*. 4:5.

[5] Smith, S. E.; Smith, F. A. (2011). Roles of Arbuscular Mycorrhizas in Plant Nutrition and Growth: New Paradigms from Cellular to Ecosystem Scales. *Annual Review of Plant Biology* 62, 227–250.

[6] Malavolta, E; Vitti, G. C.; Oliveira, S. A. (1997). *Avaliação do estado nutricional das plantas. Princípios e aplicações*. Potafos, Piracicaba.

[7] SAS-STATISTICAL ANALYSES SYSTEM. (2003). Version Release 9.0 for Windows. Cary: (CD-ROM).

[8] Neto, J. F.; Costa, C. H.; Castro, G. S. (2012). Ecofisiologia do amendoim. *Scientia Agraria Paranaensis* 11, 1–13.

[9] Hippler, F. W. R.; Moreira, M. (2013). Dependência micorrízica do amendoinzeiro sob doses de fósforo. *Bragantia* 72, 184–191.

[10] Folli-Pereira, M. D.A S. et al. (2012). Arbuscular mycorrhiza and plant tolerance to stress. *Revista Brasileira de Ciência do Solo* 36, 1663–1679.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

VALIDAÇÃO DE MÉTODOS ECOGRÁFICOS NO ESTUDO DA ARQUITETURA DO MÚSCULO MASSÉTER COM RECURSO A ANÁLISE GRÁFICA DE BLAND-ALTMAN E COEFICIENTE DE CORRELAÇÃO DE CONCORDÂNCIA

Alexandra Andé[1], João Paulo de Figueiredo [2], Luís Camilo[1], Vanessa Domingues[1]

[1] Departamento de Imagem Médica e Radioterapia, Escola Superior de Saúde de Coimbra (ESTeSC), Instituto Politécnico de Coimbra

[2] –Departamento de Ciências Complementares (Estatística e Epidemiologia), Escola Superior de Saúde de Coimbra (ESTeSC), Instituto Politécnico de Coimbra

RESUMO

Introdução: O processo da mastigação é uma das mais importantes funções do sistema estomatognático e é por isso sistematicamente estudada e referenciada. Para um adequado funcionamento do processo mastigatório é necessária toda uma harmonia neuromuscular entre um complexo conjunto de estruturas musculares, ligamentares e ósseas que são controladas pelo sistema nervoso central. Para compreender a importância desta estrutura realizam-se estudos de ecografia como exame de primeira linha para estudos dinâmicos com vista à obtenção de imagem com o objetivo de relacionar variações morfológicas faciais em indivíduos normais e definir padrões de normalização para futuras comparações. **Objetivo**: Avaliação dos métodos para aferição da espessura do maxilar (músculo masséter) com ecógrafo da marca GE Logic e por software Image J v.47.0. Material e Métodos: O estudo foi desenvolvido quer no Laboratório de Imagem Médica e Radioterapia quer num laboratório de Medicina Dentária (Coimbra). Quanto à natureza do estudo, este classificou-se do tipo observacional, analítico e de coorte transversal. **Resultados**: quando comparado os valores estimados em cada porção do músculo maxilar, quer por ecografia quer pelo software Image J (v.47), com recurso a estatística do teste t-Student, esses demonstraram, numa primeira fase do estudo um padrão médio não diferenciador ($p>0,05$). Submetemos as mesmas avaliações analíticas (imagens) ao Coeficiente Pearson, onde os resultados apresentaram, uma elevada magnitude [$r >0,8$; $p<0,001$] na maioria das comparações. Posteriormente, com recurso ao Coeficiente de Correlação de Concordância (CCC) viemos a verificar que os dois métodos de medição de imagem para aferir a espessura do músculo demonstraram, na maioria das regiões, valores de concordância bastante elevados (CCC > 0,8).

Palavras e frases chave: Ecografia, Músculo Masséter, Reprodutibilidade, Bland-Altman, Coeficiente de Correlação de Concordância.

1. INTRODUÇÃO

A hipertrofia do músculo masséter (MM), normalmente bilateral, altera vincadamente os contornos faciais através da exacerbada proeminência do ângulo mandibular e gera severo desconforto nos indivíduos. Em situações mais graves pode mesmo existir comprometimento da função muscular, levando ao aparecimento de situações clínicas como, a protusão mandibular (deslocamento frontal do maxilar), o trismo, que se refere à severa dificuldade ou impossibilidade total de uma adequada abertura da boca na amplitude desejada [(1,2)]. Este pode ser causado por anquilose óssea mandibular, comprometimento de tecidos moles por fibrose ou neoplasia do sistema nervoso periférico ou do central (através da obstrução da passagem do sinal nervoso) e o bruxismo, que consiste numa atividade parafuncional (diária ou noturna) que se traduz no ranger ou apertar dos dentes inconscientemente, levando a cefaleias frequentes e a médio-longo prazo, a um severo desgaste ortodôntico [(3,4)]. Para a realização destes estudos a ecografia tem sido

amplamente utilizada como exame de primeira linha para obtenção de imagem com o objetivo de relacionar variações morfológicas faciais em indivíduos normais e definir padrões de normalização para futuras comparações diagnósticas [6]. É um método de aquisição de imagem que provou ser capaz de fornecer informações sobre as alterações estruturais dos músculos e estudos recentes utilizaram a ecografia com a finalidade de medir a espessura dos músculos da cabeça e pescoço e correlacionar esses dados com a dor à palpação muscular, a disfunção temporomandibular (DTM), a morfologia facial e a força de oclusão dentária, de maneira a conseguir um diagnóstico clínico mais célere e eficaz [6,7,8]. Apresenta também vantagens consideráveis sobre outras modalidades de imagem na avaliação imagiológica, como a tomografia computadorizada (TC) e ressonância magnética (RM), o que a torna numa técnica mais apropriada na realização de estudos em larga escala: é uma técnica simples, rápida, de baixo custo, não invasiva e que não apresenta efeitos biológicos cumulativos conhecidos [6,8]. Contudo, a afirmação da ecografia como técnica reprodutível na avaliação do músculo masséter ainda é incerta e os índices de confiabilidade encontrados na literatura são bastante variáveis [6,8]. Tendo em conta a importância do músculo masséter na capacidade funcional de todo o sistema estomatognático e a sua importante influência em toda a estética craniofacial geral, este estudo propõe-se à caracterização arquitetónica e anatómica na população selecionada para análise estabelecendo métodos (protocolos) de diagnóstico válidos.

2. MATERIAL E MÉTODOS

A amostra deste estudo classificou-se do tipo não probabilístico e quanto à técnica esta foi de conveniência. O número amostral foi de 59 participantes, de ambos os géneros com idades compreendidas entre os 15-57 anos. A população tratou-se de indivíduos sem patologias (n=45) e com patologia (n=14). Em relação aos critérios de exclusão, foram impedidos de participar os indivíduos que apresentassem patologia do sistema músculo-esquelético, fraturas da mandíbula e antecedentes de cirurgia.

O estudo classificou-se do tipo observacional (analítico) e de coorte transversal. Os locais para a recolha de imagens foi efetuada no Laboratório J. J. Pedroso Lima da Escola Superior de Tecnologia da Saúde de Coimbra (ESTeSC) e num consultório de medicina dentária, onde se efetuou a avaliação do MM e da sua relação com a morfologia facial. Para a medição das imagens ecográficas adquiridas foi, por questões práticas e de conveniência, utilizado o software de análise científica ImageJ v.47.0. Foi posteriormente feita a comparação estatística para a correspondência entre medições efetuadas no equipamento ecográfico e aquelas obtidas no ImageJ para confirmar a confiabilidade [5] e a sobreposição dos valores analíticos obtidos. Testes estatísticos aplicados no estudo foram teste t-Student amostras emparelhadas, coeficiente de correlação linear de pearson, coeficiente de correlação de concordância e análise de bland-altman [11, 12].

3. RESULTADOS

Procuramos avaliar se existiam diferenças médias entre os equipamentos para a mesma região avaliada do MM. Para essa avaliação aplicou-se o teste t para amostras emparelhadas (tabela 1). Procuramos também avaliar o grau de concordância dos valores médios estimados pelos dois tipos de equipamentos para a mesma porção muscular. Para tal recorremos aos testes: Coeficiente de Correlação de Pearson (CCP) e ao Coeficiente de Correlação de Concordância (CCC). As fórmulas aplicadas para a estimação desses mesmos coeficientes passamos a identifica-las:

$$r_{x,y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Legenda: Coeficiente de Correlação Linear de Pearson

$$CCC = 1 - \frac{E[(y-x)^2]}{E[(y-x)^2|\rho = 0]} = 1 - \frac{\left(\mu_y - \mu_x\right)^2 + \sigma_y^2 + \sigma_x^2 - 2\rho\sigma_y\sigma_x}{\left(\mu_y - \mu_x\right)^2 + \sigma_y^2 + \sigma_x^2} = \frac{2\rho\sigma_y\sigma_x}{\left(\mu_y - \mu_x\right)^2 + \sigma_y^2 + \sigma_x^2} =$$

$$= \rho\left(\frac{2}{\frac{\left(\mu_y - \mu_x\right)^2}{\sigma_y\sigma_x} + \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y}}\right) = \rho\left(\frac{2}{v^2 + \omega + \frac{1}{\omega}}\right) = \rho\chi_a m\ que\ \chi_a = \frac{2}{v^2 + \omega + \frac{1}{\omega}}\ onde\ \omega = \frac{\sigma_y}{\sigma_x}$$

Legenda: Coeficiente de Correlação de Concordância

Com a aplicação do teste para amostras emparelhadas podemos verificar a ausência de diferenças médias significativas entre os valores estimados pelo *Ecógrafo* e pelo *Image J* para exatamente a mesma

porção do músculo em estudo (p > 0,05). O que nos permite afirmar que os dois métodos revelam ser homogéneos na quantificação dos valores médios da espessura do MM (tabela 1). Como podemos observar na tabela 1 os valores estimados pela estatística do CCC, na sua maioria, demonstraram muito boas concordâncias entre os métodos face às diferentes amostras do MM à exceção da medição em MMDO3 (CCC=0,639). Para complementar a análise de concordância recorreu-se à representação gráfica, em pares, pelos "*Diagramas de Bland-Altman*", que projeta no eixo das ordenadas a diferença absoluta das medidas de cada ponto, e, nas abscissas, sua média. Passamos a apresentar alguns diagramas que resultaram de uma muito boa concordância por CCC.

Medições	Teste t-Student				Coeficiente de Pearson		CCC	
	M	DP	DM	*p*	*r*	*p*	CCC	I.C. 95% (CCC)
EcógrafoMMEO1	0,889	0,164	0,007	0,654	0,891	<0,001	0,8902	0,771-0,949
ImageJMMEO1	0,883	0,159						
EcógrafoMMEO2	0,889	0,155	-0,001	0,967	0,882	<0,001	0,8822	0,756-0,945
ImageJMMEO2	0,889	0,154						
EcógrafoMMEO3	0,872	0,153	-0,011	0,405	0,909	<0,001	0,9065	0,804-0,957
ImageJMMEO3	0,883	0,156						
EcógrafoMMEODF1	1,064	0,192	-0,005	0,706	0,952	<0,001	0,9510	0,895-0,978
ImageJMMEODF1	1,068	0,200						
EcógrafoMMEODF2	1,077	0,186	0,002	0,881	0,963	<0,001	0,9604	0,916-0,982
ImageJMMEODF2	1,075	0,200						
EcógrafoMMEODF3	1,065	0,190	0,014	0,215	0,955	<0,001	0,9517	0,896-0,978
ImageJMMEODF3	1,051	0,185						
EcógrafoMMDO1	0,878	0,116	0,019	0,232	0,790	<0,001	0,776	0,561-0,891
ImageJMMDO1	0,859	0,129						
EcógrafoMMDO2	0,888	0,105	0,021	0,199	0,773	<0,001	0,7500	0,531-0,875
ImageJMMDO2	0,868	0,124						
EcógrafoMMDO3	0,874	0,102	0,011	0,568	0,662	<0,001	0,6394	0,362-0,813
ImageJMMDO3	0,863	0,130						
EcógrafoMMDODF1	1,051	0,160	0,005	0,765	0,882	<0,001	0,8749	0,747-0,940
ImageJMMDODF1	1,046	0,180						
EcógrafoMMDODF2	1,066	0,153	0,007	0,666	0,876	<0,001	0,8642	0,730-0,934
ImageJMMDODF2	1,058	0,178						
EcógrafoMMDODF3	1,078	0,154	0,012	0,507	0,877	<0,001	0,8544	0,721-0,927
ImageJMMDODF3	1,065	0,191						

Tabela 1: Análise de concordância das medidas estimadas por ecografia entre os dois métodos de aquisição; Legenda: M = Média; DP= Desvio Padrão; DM= Diferença Média.

Com recurso aos diagramas de "Bland-Altman" pudemos constatar que a maioria dos valores estimados pelos dois métodos de medição revelaram estar dentro da região considerada concordante (DP: -1,96; DP: +1,96) (Gráfico 1 e Gráfico 2).

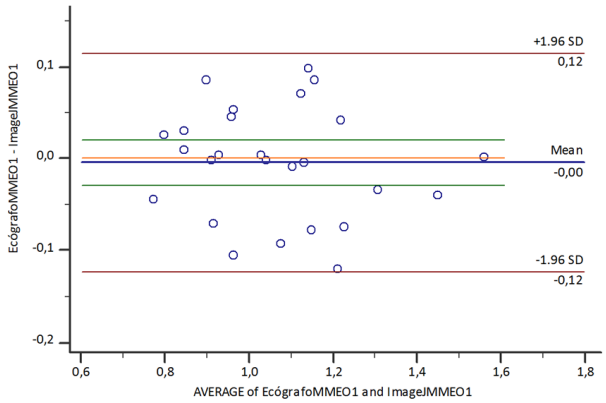


Gráfico 1

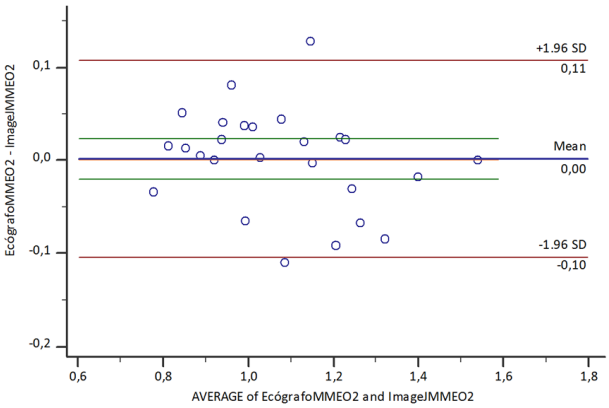


Gráfico 2.

4. DISCUSSÃO e CONCLUSÕES

Após aquisição das imagens e feita *a posteriori* as respetivas dimensões verificou-se que os resultados obtidos mostraram uma elevada confiabilidade no que se refere à avaliação das dimensões do músculo em MM. Concluímos que ambos os métodos de avaliação são fiáveis no que se refere à medição e avaliação de estruturas anatómicas.

Para a obtenção de comparações mais fidedignas, o próximo passo para uma melhor verificação destes dados seria a execução de cefalometria para medição de eixos faciais, de modo a estabelecer uma correspondência analítica entre determinado intervalo de espessura muscular e as medições dos eixos faciais que são responsáveis pelo tipo de formato facial [9, 10].

Os erros que possam estar presentes nos dados analíticos deste estudo prendem-se com a inevitável variabilidade interoperador inerente à execução de aquisições e medições ecográficas e ao facto de a obtenção de imagem no MM em relaxamento tem um maior erro associado do que em oclusão dentária forçada devido a estar mais suscetível à compressão da sonda [9].

Referências

[1]. Singh, Sourav; Shivamurthy; Varghese, D. Surgical management of masseteric hypertrophy and mandibular retrognathism. *Natl. J. Maxillofac. Surg.* 96–99 (2011).

[2]. Kebede, B. & Megersa, S. Idiopathic Masseter Muscle Hypertrophy, Case Report. *Ethiop J Heal. Sci.* **21,** 209–212 (2011).

[3]. Gonçalves, M. Prevalência e caracterização do trismo em pacientes tratados por câncer de cabeça e pescoço Prevalência e caracterização do trismo em pacientes tratados por câncer de cabeça e pescoço. (University of São Paulo, 2014).

[4]. Paesani, D. A. *et al.* Bruxism: Theory and Practic. *Quintessence Publ.* (2010). doi:Daniel A. Paesani

[5]. Kiliaridis, Stavros; Georgiakaki, Ioanna; Katsaros, C. Masseter muscle thickness and max¡llary dental arch width. *Eur. J. Orthod.* **25,** 259–263 (2003).

[6]. Mangilli, L. D., Sassi, F. C., Sernik, R. A., Tanaka, C. & Andrade, C. R. F. de. Caracterização eletromiográfica e ultrassonográfica da função mastigatória em indivíduos com oclusão normal. *J. Soc. Bras. Fonoaudiol.* **24,** 211–217 (2012).

[7]. Sassi, F. C., Mangilli, L. D., De Queiroz, D. P., Salomone, R. & De Andrade, C. R. F. Avaliação eletromiográfica e ultrassonográfica do músculo masseter em indivíduos com paralisia facial periférica unilateral. *Int. Arch. Otorhinolaryngol.* **15,** 478–485 (2011).

[8]. Bertram, S., Brandlmaier, I., Rudisch, A., Bodner, G. & Emshoff, R. Cross-sectional characteristics of the masseter muscle: an ultrasonographic study. *Int. J. Oral Maxillofac. Surg.* **32,** 64–8 (2003).

[9]. Kiliaridis, S. & Kalebo, P. Masseter Muscle Thickness Measured by Ultrasonography and its Relation to Facial Morphology. *J. Dent. Res.* **70,** 1262–1265 (1991).

[10]. Şatiroğlu, F., Arun, T. & Işik, F. Comparative data on facial morphology and muscle thickness using ultrasonography. *Eur. J. Orthod.* **27,** 562–567 (2005).

[11]. Miot, H.A. Análise de Concordância em estudos clínicos e experimentais. J Vasc Bras. 2016 Abr.-Jun.; 15(2):89-92

[12]. McBride, G.B. A Proposal for Strength-of-Agreement Criteria for Lin's Concordance Correlation Coefficient. National Institute of Water & Atmospheric Research, Ldt. New Zealand. May2005.

III Luso-Galician Meeting on Biometry
Aveiro, 28 - 30 June 2018

STRUCTURAL EQUATIONS MODEL OF A QUESTIONNAIRE ON THE PATIENT SAFETY CULTURE IN PORTUGUESE PRIMARY CARE

Carina Silva[1,2], Margarida Eiras[1,3]

[1]Escola Superior de Tecnologia da Saúde de Lisboa, IPL
[2]Centro de Estatística e Aplicações, Universidade de Lisboa
[3]Centro de Investigação em Saúde Pública, Universidade Nova

ABSTRACT

Patient safety in Primary Health Care has been studied in recent years along with other healthcare areas. Assessing patient safety culture is a strategic priority worldwide and Portugal is no exception. It is the objective of this work to translate, adapt, validate and analyze the reliability of the Medical Office Survey on Patient Safety Culture. The methodology adopted focused on transcultural translation, adaptation using the Translation Guidelines for the AHRQ Surveys on Patient Safety Culture and validation and analysis of the reliability of the instrument was performed using a Structural Equation Model (SEM) and Confirmatory Factor Analysis (CFA) analysis, checking and modifying its model by Wald and Lagrange indicators to obtain the most adjusted model to the theoretical and goodness criteria.

Keywords and key sentences: Patient safety, primary care, validity, reliability, Structural Equation Model, Confirmatory Factor Analysis.

1. INTRODUCTION

Patient Safety is considered to reduce the risk of unnecessary damage related to health care to an acceptable minimum[1]. The safety culture of an organization is the product of individual and group values, attitudes, perceptual capacities, and behavioral patterns that determine the commitment to an organization's health and safety management and its style and proficiency[2].

In recent years, in Portugal, patient safety has been a topic of study especially in hospital care[8]. Recently there has been a growing interest in the study of this subject also in Primary Health Care (PHC). Evaluating and monitoring the patient's safety culture is considered to be the first step towards building and consolidating an open and fair culture that allows all professionals to become involved in improving the organization's culture[3]. In this sense, several instruments were developed and the AHRQ (Agency for Healthcare Research and Quality) designed a questionnaire to evaluate the safety culture of the patient in the primary

health care, called the Medical Office Survey on Patient Safety Culture (MOSPSC).

In Portugal, following the publication of the Strategy for Quality in Health[4] and more recently of the National Plan for Patient Safety[5], the Department of Quality in Health published a guideline for promoting the evaluation of the patient's safety culture in primary health care. From 2015 onwards, and every two years, the evaluation of the patient's safety culture in primary health care should be carried out in the Health Center Group, through a questionnaire to be filled out by all professionals and collaborators.

Guidelines for translating patient safety culture questionnaires published by the Agency for Healthcare Research and Quality (AHRQ) have been adopted. After, the questionnaires were distributed to 62 enrolled Health Units and answered by 54, corresponding to 4596 respondents. After an analysis of missing values it was decided to remove all the questionnaires with missing answers, getting a final sample of 2379.

The questionnaire contains 38 items that assess 10 dimensions of the patient safety culture (Table 1) and all were assessed with a five-point Likert scale, ranging from *strongly disagree* or *never* to *strongly agree* or *always*.

Dimensions	**N. Items**
Teamwork (D1)	4
Follow-up of the patient (D2)	4
Organization learning (D3)	3
General perception of quality and patient safety (D4)	4
Training of the professionals (D5)	3
Support by top management (D6)	4
Communication about errors (D7)	4
Openness in communication (D8)	4
Administrative procedures and procedures standardization (D9)	4
Pressure and work pace (D10)	4

Table 1: Dimensions (D) and their respective number of items.

A reliability analysis was conducted using Cronbach's alpha. However, the validity of this measure has been questioned and several authors have suggested alternative measures. In this study we also used the average inter-item correlation (AIIC).

Validity refers to how well the instrument measures what it is intended to quantify. Construct validity is considered the most valuable indicator[7]. Composite scores and inter-correlations allow us to analyse construct validity. The construct validity of each safety culture dimension would be reflected in composite scores moderately related to one another.

We used confirmatory factor analysis (CFA) to compare the Portuguese sample factor structure to the factor structure reported for the original HSOPSC[8]. We used chi-square divided by degrees of freedom, where the model fit is considered good if the quotient is less than 2. Less than 5 is acceptable and values greater than 5 are unacceptable. We also used the goodness-of-fit index (GFI), which accounts for the proportion of observed covariance between the manifest variables (items), explained by the fitted model (a concept similar to the coefficient of determination in linear regression). Generally GFI values between 0.9 and 0.95 indicate good fit and GFI values above 0.95 indicate a very good fit. Bentler's Comparative Fit Index (CFI) was used to correct the under-estimation that can occur when samples are

small. CFI is independent from the sample size. Values between 0.9 and 0.95 indicate good fit and values equal to or above 0.95 indicate a very good fit. The Tucker-Lewis index (TLI) varies between 0 and 1, values close to 1 indicate a good fit. Parsimony GPI (PGFI) is obtained to compensate for the "artificial" improvement in the model, which is achieved simply by adding more parameters, i.e., a more complex model may have better fit than a simpler model (parsimonious). Values between 0.6 and 0.8 indicate a reasonable fit and values above 0.8 a good fit. The index Root Mean Square Error of Approximation (RMSEA) was used to adjust the model simply by adding more parameters. Empirical studies suggest that the model fit is considered good for values ranging between 0.05 and 0.08 and very good for values less than 0.05.

Statistical analysis were conducted using `R` software.

2. RESULTS

Reliability analysis using Cronbach's alpha were performed on 10 dimensions to ensure that individuals were responding consistently to items (Table 2). There are two dimensions with Cronbach's Alpha lower than 0.7 (D5 and D9), however very close to 0.7. The highest value was achieved by D1. The instrument as a whole achieved a high Cronbach's value, reviling a good internal consistency. Other internal consistency measure is the average inter-item correlation (Table 2). This measure evaluate how items within a dimension correlate, that is there is evidence that the items are measuring the same underlying dimension. A rule-of-thumb is that AIIC should be at least 0.3 and all dimensions had values of AICC higher than 0.3.

Dimensions	**Cronbach's Alpha**	**AIIC**
Teamwork (D1)	0.832	0.556
Follow-up of the patient (D2)	0.759	0.442
Organization learning (D3)	0.811	0.593
General perception of quality and patient safety (D4)	0.724	0.413
Training of the professionals (D5)	0.696	0.432
Support by top management (D6)	0.727	0.400
Communication about errors (D7)	0.787	0.482
Openness in communication (D8)	0.739	0.419
Administrative procedures and procedures standardization (D9)	0.679	0.348
Pressure and work pace (D10)	0.749	0.413
Total	0.934	

Table 2: Internal consistency statistics.

The model need to be validated in a different sample from that where the model was adjusted. A common model validation is the cross-validation strategy when the sample size is large. In this case, considering the rule-of-thumb of 10 respondents for each item (38), it is needed at least 380 respondents where it was far achieved with a sample size of 2379. So, two thirds of the whole sample, randomly selected, is used to adjust the model (n=1808) and the remainder (n=571) is used to evaluate the model's invariance. If the fitted model in the first sample provides a good fit in the second, then we can assume that the model is invariant in the two samples and the model is valid for the population.

Firstly it was analyzed the theoretical model developed by [7] on Portuguese Primary Care, ie, considering the 10 dimensions. Normality assumptions were verified. Considering the 10

dimensions the covariance matrix of latent variables were not positive definite. This is due to existence of high correlate dimensions. Analysing the correlation matrix it was observed that dimension 3 was highly correlated with dimensions 4, 7 and 9. Considering that it was removed dimension 3 and was considering the remaining dimensions it was obtained the model in Figure 1.

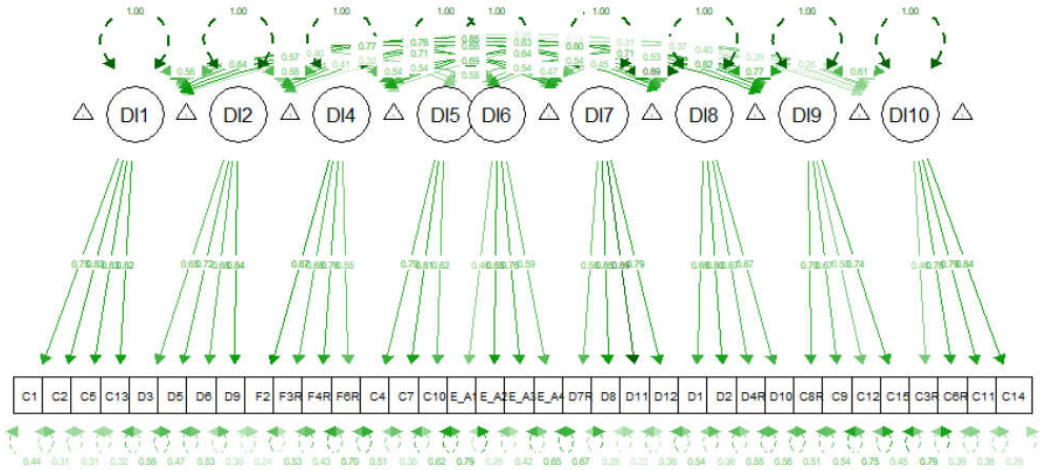


Figure 1: Confirmatory Factor Analysis diagram.

The analysis indicated that the model has a very good overall fit considering the indexes: CFI=0.97; GFI=0.98; TLI=0.97. Considering the indexes PGFI=0.7; RMSEA=0.08 the model has a good overall fit. But considering $\chi^2/df = 12.43$, pvalue¡0.001 the model has a poor overall fit, however χ^2 is highly influenced by the sample size. It was conducted the same analysis but considering the other subsample (n=571). Considering the same model the covariance matrix of the latent variables were not positive definite. Considering both samples there is no consistency in the results.

3. CONCLUSIONS

The instrument showed a satisfactory reliability on the 10 dimensions and considering the questionnaire as a whole. The confirmatory factor analysis on the first subsample had good results when the third dimension was removed. However, considering the same model on the second subsample the covariance matrix of the latent variables were not positive definite. Those contradictory results lead us to consider that this instrument needs to be revaluated considering other models.

ACKNOWLEDGMENT

This work is partially financed by national funds through FCT Fundação para a Ciência e a Tecnologia under the project UID/MAT/00006/2013.

References

[1] Direção-Geral da Saúde. (2011). *Estrutura Conceptual da Classificação Internacional sobre Segurança do Doente*. Relatório Técnico Final. Tradução realizada pela Divisão de Segurança do Doente, Departamento da Qualidade na Saúde. Lisboa.

[2] National Patient Safety Foundation (2001). *Forum of end Stage Renal Disease Networks, National Patient Safety Foundation, Renal Physicians Association, Renal Physicians Association.* National ESRD Patient Safety Initiative: Phase II Report. Chicago.

[3] National Patient Safety Agency. (2006). *Seven steps to patient safety for primary care*. London.

[4] Ministério da Saúde. (2015). *Estratégia Nacional para a Qualidade na Saúde 2015-2020*. Despacho n.º 5613/2015.

[5] Ministério da Saúde. (2016). *Plano Nacional para a Seguranca dos Doentes 2015-2020*. Gabinete do Secretario de Estado Adjunto do Ministro da Saude. Despacho nº 1400-A/2015

[6] Sorra, J. and Dyer, N. (2010). Multilevel psychometric properties of the AHRQ hospital survey on patient safety culture. *BMC Health Services Research*, 10, p. 199.

[7] Sorra, J. and Nieva, F. (2004). *Hospital survey on patient safety culture: 2004 report*. AHRQ Publication no. 04-0-041, Agency for Healthcare Research and Quality, Rockville, MD.

[8] Eiras, M., Escoval, A., Grillo, I.M. and Silva-Fortes, S. (2014). The hospital survey on patient safety culture in Portuguese hospitals. *International Journal of Health Care Quality Assurance*, 27(2):111–122.

III Encontro Luso-Galaico de Biometria
Aveiro, 28 - 30 de junho de 2018

MODELAÇÃO CONJUNTA DE DADOS LONGITUDINAIS E DADOS DE SOBREVIVÊNCIA NA PRESENÇA DE RISCOS COMPETITIVOS

Laetitia Teixeira[1], Inês Sousa[2], Anabela Rodrigues[3] e Denisa Mendonça[4]

[1]Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto & CINTESIS, Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto & EPIUnit, Instituto de Saúde Pública, Universidade do Porto
[2]Departamento de Matemática e Aplicações, Universidade do Minho,& Centro de Biologia Molecular e Ambiental – CBMA
[3]Departamento de Nefrologia, Centro Hospital do Porto – Hospital Geral de Santo António & Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto
[4]Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto & EPIUnit, Instituto de Saúde Pública, Universidade do Porto

RESUMO

A análise conjunta de dados longitudinais e de dados de sobrevivência na presença de riscos competitivos tem vindo a verificar uma atenção especial, resultando no aumento de abordagens propostas para este tipo de dados. No entanto, a sua disseminação e utilização ainda é escassa, nomeadamente a sua aplicação em estudos clínicos.

A existência de diferentes abordagens no contexto da análise conjunta de dados longitudinais e de dados de sobrevivência na presença de riscos competitivos requer uma adequada interpretação dos resultados, tornando essencial uma análise comparativa destes modelos.

O presente estudo tem como principal objectivo a modelação conjunta de dados longitudinais e de sobrevivência na presença de riscos competitivos, discutindo diferentes parametrizações de implementações sistemáticas destes modelos no software livre R.

Para demonstrar a relevância da análise conjunta de dados longitudinais e de dados de sobrevivência na presença de riscos competitivos em investigação clínica, os diferentes modelos serão aplicados a dados no contexto da diálise peritoneal. Serão ainda comparados os resultados da modelação conjunta com os dados obtidos considerando abordagens separadas (dados longitudinais e dados de sobrevivência).

Palavras e frases chave: Modelação conjunta; dados longitudinais; dados de sobrevivência; riscos competitivos; diálise peritoneal.

III Encontro Luso-Galaico de Biometria

Aveiro, 28 - 30 de junho de 2018

FUNÇÃO DE LIGAÇÃO DE CAUCHY PARA AVALIAÇÃO DE P_{50} de LONGEVIDADE DE SEMENTES DE SOJA

Amanda Rithieli Pereira Dos Santos[1], Rute Quelvia de Faria[1], Deoclecio Jardim Amorim[1], Edvaldo Aparecido Amaral da Silva[1], Maria Márcia Pereira Sartori[1]

[1] Universidade Estadual Paulista "Júlio de Mesquita Filho"; Faculdade de Ciências Agronômicas, Departamento de Melhoramento e Produção Vegetal, Botucatu, Brasil.

RESUMO

O estudo do momento em que um lote de sementes esta em 50% de sua viabilidade (P_{50}) é uma importante ferramenta utilizada para manutenção dos bancos de recursos genéticos. Esse valor normalmente é definido usando a função de ligação de Probit, o que não é eficaz para algumas espécies. Assim este trabalho objetivou estudar a viabilidade do uso da função de ligação de Cauchy em comparação com Probit para estimação de P_{50} no estudo de longevidade de sementes de soja. Utilizou-se 3 lotes de sementes, foram avaliados a germinação e longevidade, as curvas de sobrevivência dos dados de Longevidade obtidos através da avaliação da germinação no tempo. Após realizados as transformações dos dados com as Funções de Probit e Cauchy, realizou-se o ajuste linear e determinação do P_{50}. Conclui-se que o uso da função de Cauchy é uma alternativa viável para a estimação do P_{50} de dados de longevidade de sementes.

Palavras e frases chave: *Glycine max*, Funções simétricas, estimação.

1. INTRODUÇÃO

A manutenção dos bancos de germoplasma e recursos genéticos são realizados através do estudo de longevidade de sementes. O estudo da longevidade de sementes é a parte da ciência responsável por avaliar o período máximo de tempo que uma semente pode permanecer com viabilidade em função do tempo de armazenamento, quando realizado em condições ambientais favoráveis [6].

Para a previsão da Longevidade de sementes foram desenvolvidas as equações básicas de viabilidade [7] que posteriormente foram melhoradas [3]. Nessas publicações os autores propõem o uso da função de ligação de Probit para estimação do período de 50% de viabilidade do lote avaliado (P_{50}). Esse valor é encontrado através dos parâmetros ajustados da regressão, e possuem pré-requisitos para seu uso, como normalidade dos dados [3].

Autores como Schneider [8] e Hill [4] não conseguiram estimar a longevidade pela função de probit, fazendo-se necessário o estudo de outras funções de ligação aplicado a

longevidade de sementes. Bonat [2] em seu estudo sobre regressões explica que para áreas como à agronômica, existem uma gama de funções que podem ser estudadas, o que deve ser feito em acordo com a necessidade e característica dos dados, neste caso binário.

Este trabalho teve como objetivo estudar a viabilidade do uso da função de ligação de Cauchy em comparação com a função de ligação de Probit para estimar o P_{50} no estudo de longevidade de sementes de soja.

2. MATERIAL E MÉTODOS

Os lotes de sementes de soja foram adquiridos em diferentes regiões geográficas do Brasil, neste trabalho foram analisados 3 lotes, respectivamente das regiões: Sul, Centro e Norte. As amostras obtidas foram submetidas aos testes de Germinação e posteriormente as sementes foram preparadas para a condução da avaliação de longevidade. O teste de Germinação foi conduzido conforme Brasil [1], utilizou-se 6 repetições de 50 sementes. Para a longevidade as sementes que foram colocadas em caixas plásticas tipo gerbox contendo aproximadamente 300 sementes cada, sem sobreposição, e foram armazenadas em ambiente controlado com 75% de umidade relativa do ar (Sal utilizado: NaCl) a 35°C pelo número de dias necessário, ou seja, enquanto houvesse viabilidade [5]. As sementes foram avaliadas semanalmente quanto à germinação e ao teor de água até a estabilização.

Os dados de longevidade (curvas de sobrevivência) foram transformados pelas funções de ligação descritas na Tabela 01.

Função de Ligação	Modelo	Equação
Probit	$F(x) = \Phi^{-1}$	1
Cauchy	$F(x) = \tan\left(\pi\left(x - \frac{1}{2}\right)\right)$	2

Tabela 01. Modelos das funções de Ligação de Probit e Cauchy.

Para a função de Cauchy os ajustes foram realizados utilizando 6 repetições, no entanto, para a realização dos ajustes das regressões lineares a partir dos dados transformados em Probit foram utilizadas médias conforme Ellis e Roberts [3]. Foi realizada correção dos dados antes da transformação em Probit para garantir a simetria. A multiplicação realizada foi de 0,9999999713 em todos os pontos, exceto 0 que teve o valor trocado por 2,87 x 10^{-07}, valor definido a partir de simulações realizadas em pré-testes. O mesmo processo foi realizado para a função de Cauchy, o valor definido foi 0,95 e 0,05 (valor esse substituído no zero).

Os ajustes das regressões foram avaliados em função da proximidade do intervalo esperado (dados experimentais) para o valor de P_{50} com relação ao estimado para cada função de ligação avaliada. As análises dos dados foram realizadas com o programa R 3.4.1.

3. RESULTADOS E DISCUSSÃO

Na imagem 1(A) são apresentados os dados de viabilidade em função do tempo de armazenamento de cada lote. Observa-se que cada lote apresentou diferentes longevidades, sendo o lote 3 o com maior período de viabilidade. Nas imagens 1(B, C e D) são apresentados os ajustes lineares de cada lote, sendo os parâmetros dos ajustes apresentados na tabela 2.

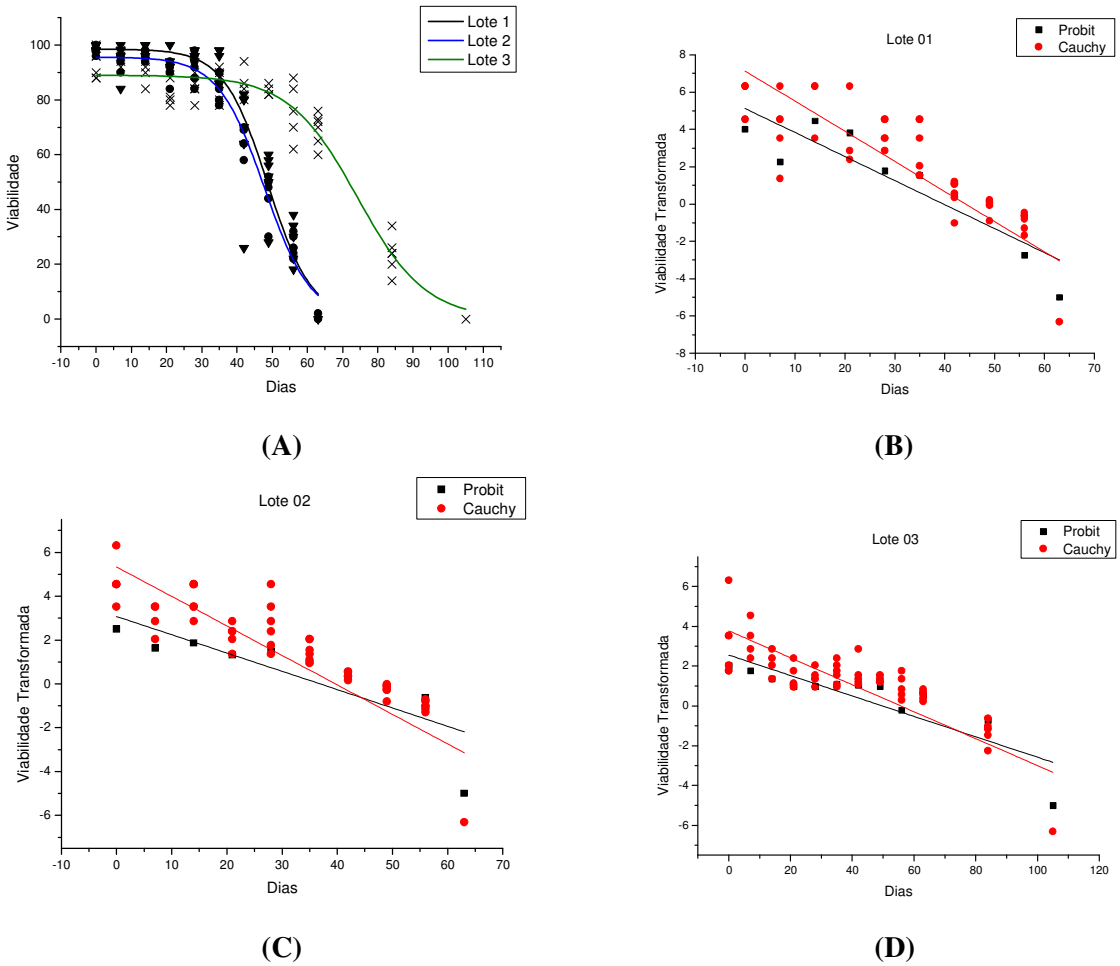


Imagem 1. (A) Dados de Viabilidades de cada lote em função do tempo de armazenamento. Ajustes lineares para as funções de Probit e Cauchy para (B) lote 1, (C) lote 2 e (D) lote 3. Brasil (2018).

Lote	Região	Intervalo esperado	Probit			Cauchy		
			Intercept	Slope	P 50	Intercept	Slope	P 50
1	Sul	42 a 49	5,12	0,13	39,8	7,13	0,16	44,2
2	Centro	42 a 49	3,08	0,08	36,8	5,34	0,13	39,7
3	Norte	63 a 84	2,55	0,05	49,7	3,77	0,07	55,7

Tabela 02. Resultados dos dados de cada lote para: intervalo esperado, intercept e slope dos ajustes lineares e P_{50} obtidos para Função de Probit e Cauchy, respectivamente.

Para todos os lotes avaliados foram caracterizados os valores de P_{50} experimentais, (Tabela 2), nota-se que para os lotes 1 e 2 os valores experimentais de P_{50} necessariamente devem estar contidos no intervalo de 42 a 49 dias, já para lote 3 esse intervalo é entre 63 e 84 dias.

O valor estimado de P_{50} do lote 1 foi de 39,8 e 44,2 dias para Probit e Cauchy, respectivamente. Nota-se que o valor de P_{50} estimado pelo Probit, subestima esse parâmetro em 2,2 dias, no entanto, o valor resultante de Cauchy encontra-se dentro do intervalo de interesse, indicando a confiabilidade da estimação. Para os lotes 2 e 3 observamos que os valores obtidos não estão contidos no intervalo de interesse, entretanto ao avaliar-se o P_{50} é notório que os dados transformados com a função de ligação de Cauchy aproximou-se mais do intervalo desejado que os obtidos pela função de ligação de Probit, encontrando-se uma diferença de até 6 dias. Esse

contexto corrobora com Hill [4] que sugere que outras funções podem estimar a longevidade com maior robustez que Probit.

Embora a função de Cauchy possa ser utilizada para estimar parâmetros de longevidade é necessário estudo de outras funções ou correções capazes de melhorar ainda mais essa estimativa.

4. CONCLUSÕES

A utilização da função de Cauchy para a avaliação do P_{50} de dados de longevidade de sementes é viável, pois ela é capaz de estimar valores dentro do intervalo de interesse ou com maior proximidade desse intervalo. No entanto são necessários maiores estudos para determinar suas limitações e outras alternativas que englobem maiores aplicações em diversas espécies existentes.

AGRADECIMENTOS

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). À Faculdade de Ciências Agronômicas (FCA) da Universidade Estadual Paulista – Júlio de Mesquita Filho (UNESP).

Referências

[1] BRASIL. Ministério da Agricultura, Pesca e Abastecimento. Regras para análise de sementes. Secretaria Nacional de Defesa Agropecuária. Brasília, 2009.

[2] BONAT, W.H; RIBEIRO JR, P. J.; ZEVIANI, W.M.. Regression models with responses on the unity Interval: specification, estimation and comparison. Rev. Bras. Biom., São Paulo, v.30, n.4, p.415-431, 2012.

[3] ELLIS, R.H.; ROBERTS, E.H. Improved equations for the prediction of seed longevity. Annals of Botany, v.45, p.13-30, 1980.

[4] HILL, H.J.; CUNNINGHAM, J. D.; BRADFORD, K.J.; TAYLOR, A.G. Primed Lettuce Seeds Exhibit Increased Sensitivity to Moisture Content During Controlled Deterioration. HortScience October 2007 42:1436-1439.

[5] PEREIRA LIMA, J.J.; BUITINK, J.; LALANNE, D; ROSSI, R.F.; PELLETIER, S.; DA SILVA, E.A.A; et al. (2017). Molecular characterization of the acquisition of longevity during seed maturation in soybean. *PLoS ONE* 12(7): e0180282.

[6] PEREIRA NETO, LEONEL GONÇALVES. Longevidade de sementes de Astronium fraxinifolium Schott: estudos fisiológicos, bioquímicos e moleculares. 2016. 136 f. Tese (Doutorado) - Curso de Agronomia, Agricultura, Universidade Estadual Paulista - Faculdade de Ciências Agronômicas (FCA), Botucatu, 2016. <https://repositorio.unesp.br/handle/11449/148714>.

[7] ROBERTS, E.H. Storage environment and the control of viability. In: ROBERTS, E.H. (Ed.). Viability of seeds. New York: Syracuse University Press, 1972. P.14-58.

[8] SCHNEIDER, CRISTINA FERNENDA et al. Equações de longevidade para sementes de pau-marfim. Revista de Ciências Agrarias - Amazon Journal Of Agricultural And Environmental Sciences, [s.l.], v. 60, n. 1, p.53-59, 2017. Editora Cubo Multimidia.

Index

Sociedade Portuguesa de Estatística

#SGaPEIO
Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións

Parceiros e Patrocinadores

dmat universidade de aveiro departamento de matemática

FCT Fundação para a Ciência e a Tecnologia

FÁBRICA CENTRO CIÊNCIA VIVA aveiro

BIOMATH CIDMA THEMATIC LINE

CEAUL Centro de Estatística e Aplicações Universidade de Lisboa

CIDMA

CINTESIS Health.Research.

DELTA CAFÉS

EDIÇÕES SÍLABO Publicamos conhecimento

Profijardim® Construção e Manutenção de Espaços Verdes

PSE

Porto de AVEIRO

INSTITUTO NACIONAL DE ESTATÍSTICA STATISTICS PORTUGAL

Jerónimo Martins